Author:
Laimann, Jessica M

Title:
Classification and Explanation at the Crossroads of the Social and Natural Sciences

CLASSIFICATION AND EXPLANATION AT THE CROSSROADS OF THE SOCIAL AND NATURAL SCIENCES

Jessica Melanie Laimann

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Arts.

University of Bristol
Faculty of Arts
Department of Philosophy

June 2019

Word count: 77,651

ABSTRACT

This thesis looks at a number of debates relating to how practices of classification and explanation in the social sciences differ from those in the natural sciences. Against the background of an extant discussion that is fragmented and sometimes contradictory, my thesis develops the notion of *hybrid kinds* as a key unifying concept for understanding and resolving a range of debates on the relationship between the natural and social sciences. Chapter 1 sets out my account of hybrid kinds. By critically evaluating the discussion of Searle's theory of social kinds, I argue that the social kinds in question are best understood as hybrid kinds that are constituted of a base kind and a status kind linked by a relation of status conferral. Chapter 2 applies the hybrid kind account to Epstein's grounding-anchoring model of social kinds and argues that the latter is fundamentally flawed. Chapter 3 turns to the question of moral and political values in the semantics and ontology of social kinds and explores how the direct and indirect influence of such values could be justified. Chapter 4 marks the move from largely ontological considerations to exploring the role of hybrid kinds in scientific contexts. Against the widespread criticism of Hacking's claim that human kinds cannot be natural kinds, I argue that an appropriate understanding of the epistemic role of natural kinds and of the capricious nature of social feedback effects renders his claim more viable than hitherto acknowledged. Chapters 5 and 6 use the previous findings on hybrid kind ontology and epistemology to advance a central debate at the crossroads of the natural and social sciences: the discussion on biological versus social explanations of gender differences in human behaviour and psychology.

Für Sonja, Rainer und Tobias

ACKNOWLEDGEMENTS

Studying philosophy and completing a PhD at elite universities in the UK requires a good amount of funding, professional support and luck and I have much to be thankful for in all three regards. In terms of funding, I am grateful to the Arts and Humanities Research Council and the German Academic Scholarship Foundation for the financial support that made this journey possible. In terms of professional support, I want to thank my supervisors and fellow academic travellers for their invaluable advice and for making the long hours spent on this project meaningful. I would like to thank Samir Okasha for the many thorough and encouraging discussions of ideas that I imagine were often quite different from his own, John Dupre for insightful criticism of ideas that I suspect tended to align relatively well with his own, and Karim Thebault, whose eye for detail and skilled editorial advice were vital to my success in publishing the paper Capricious Kinds (Chapter 4 of this thesis) in the British Journal for the Philosophy of Science. I am very grateful to Michelle, Maxime, Katie, Joanna, Sam, Stephen, Kam, Patricia and Ellen for creating a spirit of support and collaboration in the midst of the alienating place that was Oxford, and to Snai, Lilit, Georgie, Alice, Henning, Tyrone, Zack and Rhian for making the PhD experience anything but lonely and for being passionate about things beyond philosophy. In terms of luck, I am indebted to my wonderful family, friends and other supportive people I was fortunate enough to meet along the way. Vielen Dank an meine Eltern, die mich auf fernen Reisen in die akademische Welt nie den Blick für das Wesentliche haben verlieren lassen. In addition to the friends already named above, my sincerest thanks go to Charlott, Johanna, Tristan, Claas, Sven and Marius for their overwhelming helpfulness in supporting the tedious editing process – Ostwestfalians do make the most loyal friends. Thanks also to Niko and Maria, who made philosophy interesting and academia a place where I could imagine myself, even though it wasn't for me in the end. Special thanks to Sam and the Welsh hills for showing me better places.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: DATE:

CONTENTS

INTRODUCTION

This thesis looks at a number of debates relating to issues of classification and explanation in the social and natural sciences. It investigates how – and why – practices of classification and explanation in the social sciences sometimes differ from those in the natural sciences. While these questions are neither very new nor very specific, I believe the following chapters address them in a way that is both novel and philosophically fruitful.

To understand why, it helps to consider some characteristics of the extant debate on these questions. Many philosophical discussions try to show that the social and the natural sciences are essentially not that different. These discussions typically start with models of scientific method established within the context of the natural sciences and show how they can be applied to the social sciences. While this approach has some merit it also has important limitations. Most importantly, it predisposes us to look for similarities between the natural and the social sciences and potentially encourages us to shy away from differences where we shouldn't.

If we turn to the philosophical literature that investigates the social sciences on their own merit, rather than against the standard of natural scientific methodology, different problems emerge. Although these approaches set out to embrace the idiosyncrasies of the social sciences, the debate is so pluralist and often contradictory that it becomes difficult to identify special characteristics of the social sciences that any two philosophers agree on. We will consider individual approaches in more detail in the following chapters, but here are some brief examples for the purpose of illustration. Consider the topic of classification in the social sciences, that is, social kinds. These are sometimes also referred to as "institutional kinds". Broadly speaking, social kinds are groupings of entities defined by a social property, i.e. a property which cannot exist in the absence of (human) social life.[1] Accordingly, social kinds include a diverse range of classifications such as friend, money, tenant, limited company, football game, working class, war and recession.

Some philosophers argue that social kinds are best understood through the lens of game theory, some insist they involve a distinct relation of "status conferral", and others suggest that they can only be understood in terms of two distinct metaphysical relations ("grounding" and "anchoring"). Some of these philosophers conclude that social kinds

[1] I think we can plausibly talk of animal social life and hence potentially animal social facts and social kinds. However, the discussion in this thesis focusses on the human realm.

are essentially the same as natural kinds, others insist that they are fundamentally different; some claim that social kinds are importantly normative, others deny this; and so forth. It is easy to lose one's orientation in this fragmented debate. How do these different claims relate? And more importantly, which ones – if any – are correct?

To complicate things further, there is significant overlap with the discussion on human kinds. Conceptually, human kinds can be distinguished from social kinds because the former are best understood as classifications of human individuals or human properties. Some social kinds are human kinds (tenant, friend, working class), but others are not because they are not classifications of human individuals or properties (money, recession, war). At the same time, not all human kinds are necessarily social kinds. The kind *blood type A-negative* is a classification of human individuals, but it does not make reference to a social property. Yet, as we will see in later chapters, even human kinds that are not defined in terms of social properties have an elusive tendency to become entangled in the social.

The situation is no less confusing when we turn to the topic of explanation, and the question how social and natural scientific explanations of the same target phenomenon relate. Probably the most prominent case in point – and a core case study in this thesis – is the debate between biological explanations of gender differences in psychology and behaviour on the one hand and social scientific explanations of the same target phenomena on the other. Many people who are committed to feminism and gender equality would argue that biological explanations in this context reflect and reinforce harmful gender stereotypes. Accordingly, they tend to support social scientific explanations as superior explanations of the phenomena in question. The political fault lines are clear, the causal-explanatory ones much less so. Are these two types of explanations in conflict, as some people have suggested, or are they compatible, as others insist? And how does the fact that the stakes in this debate are so expressly political impact on the question?

My thesis aims to bring some clarity into this confusion by using a key conceptual tool: the model of *hybrid kinds*. Simply speaking, a hybrid kind is a classification of entities that is associated with a social dimension or meaning. Due to the associated social meaning, the classification is leading a double life: on the one hand, it is picking out a group of not necessarily social entities (the base kind), on the other hand, it describes a group of entities with specific social properties (the status kind). What connects base kind and status kind is the fact that members of the base kind acquire the relevant social properties in virtue of being recognised as members of the kind. I argue that a central characteristic of many

social and human kinds can be understood in terms of this hybrid structure: by association with a social meaning, classifications simultaneously refer to a cluster of social properties (the status kind) and to a group of whatever underlying entities (the base kind).

Money is a good example of a social kind that can be understood as a hybrid kind. The kind consists, among other things, of metal coins and paper bills (the base kind) which have acquired the social function of being a means of exchange (the status kind). Furthermore, the reason the coins and bills function as a means of exchange is because people collective recognise them as something over and above metal and paper. In a sense, it seems that we need to recognise something as money in order for it to function as money. This claim will be spelled out in more detail Chapter 1.

The core idea is comparatively simple and has popped up in different places both in analytic social ontology and in social theory. However, I believe the idea's potential as a unifying tool for understanding and resolving a range of debates on the relationship between the natural and social sciences has been hugely underestimated. My thesis explores this uncharted territory by developing the scattered remarks on the dual nature of social kinds into a coherent account of hybrid kinds. This hybrid kind model will not only provide a powerful tool for clarifying a number of ongoing discussions on the ontology and epistemology of social and natural kinds. It will also prove a fruitful starting point for investigating the questions of causal explanation and explanatory pragmatics touched on above.

There are, of course, important limitations. For one thing, the hybrid kind model does not provide a general model of social ontology. There are, as will become clear over the course of this thesis, social kinds that are not hybrid kinds. Social classifications that lack a distinct social meaning will fall under this category. Examples include social kinds that are not widely recognised, for instance social scientific kinds which are not generally known outside a specific academic discipline (Zygmund Bauman's tourists and vagabonds[2]) and social classifications which are not considered significant in a wider social context (Sainsbury customer, people with more than three best friends).

Furthermore, my thesis does not claim that the hybrid kind model is the only or the single most fruitful way of thinking about the phenomena in question. What it does suggest is that the hybrid kind model is an invaluable tool to clarify and "streamline" a range of fragmented conceptual debates on social ontology and social explanations. Finally, the hybrid kind model and the ensuing analysis of scientific debates do not aim

[2] Bauman 1998.

to prove a universal recipe for resolving any and all discussions between social and natural scientists on matters of political concern. Given the limited number of case studies that my inquiry focusses on, this claim would be unfounded. What my thesis does hope to achieve, however, is to overcome entrenched but fruitless ways of thinking about these matters and pave the way toward novel and more fruitful ones.

The thesis proceeds as follows. Chapter 1 sets out my account of social kinds as hybrid kinds. I begin by critically evaluating Searle's account of social kinds, as well as the extant discussion of Searle's account exemplified by Amie Thomasson and Francesco Guala. I argue that two characteristics of social kinds can be distinguished in Searle's original discussion: *reflexivity* – the idea that social kinds have no instances without someone having an understanding of what they are – and *hybridity* – the idea that social kinds involve imposing a social status unto an existing classification of entities. Reflexivity has received the lion's share of attention in the extant debate on Searle's account – Thomasson endorses reflexivity, Guala rejects it. I argue that although Guala is right to reject reflexivity, critics and proponents of reflexivity alike have failed to recognise that the most valuable element of Searle's account of social kinds is the idea of hybridity. As a result, neither Guala nor Thomasson provide a convincing account of the social kinds in question. A more promising understanding needs to focus on hybridity – the idea that social kinds involve a special sort of relation by which social statuses or properties are imposed onto existing entities. On the basis of these insights, I develop my hybrid kind model, which understands the kind in question as hybrid kinds that are constituted of a base kind and a status kind linked by a relation of status conferral. I then use this model to clarify questions about the ontology and epistemology of the kinds in question.

Chapter 2 applies the hybrid kind model to another high-profile discussion in social ontology – the discussion on Epstein's grounding-anchoring model. Epstein's grounding-anchoring distinction has received much critical acclaim. Nevertheless, many critics reject the grounding-anchoring distinction and suggest that anchoring can simply be understood as a form of grounding. I argue that Epstein's grounding-anchoring model is flawed more fundamentally than his critics realise. Unwittingly, Epstein bases his model on an erroneous understanding of hybrid kinds that he then tries to extend to non-hybrid social kinds. As a result, the model is fit for neither. I conclude the chapter by providing an error theory of the intuitions motivating the grounding-anchoring model.

Chapter 3 explores the role of moral and political (so-called "non-epistemic") values in the semantics and ontology of social kinds. Some people insist social ontology has no

room for these values. Others object that non-epistemic values have a legitimate, maybe even necessary, say in how we conceptualise the social world. A prominent advocate of the later position is Sally Haslanger, who proposes that we should define social concepts in a way that best suits our legitimate purposes, which may be a question of explicitly moral or political considerations. Haslanger's proposal has been heavily criticised by Francesco Guala. Guala argues that the question how to conceptualise the social world cannot be left "up to us" and the outcome of our potentially controversial moral and political disputes. Interestingly, both Haslanger and Guala rely on semantic externalism to defend their positions. I argue that recent criticism of semantic externalism lends support to the idea that moral or political values play an important *indirect* role in scientific ontology. I then point out that several of Haslanger's examples advocate a direct rather than indirect role of these values. Using the hybrid model of social kinds developed in Chapter 1, I clarify how exactly the cases differ, and explore how the direct influence of moral and political values could be justified.

Chapter 4 marks the move from largely ontological considerations to exploring the role of hybrid kinds in scientific contexts. I consider the long-standing debate on Hacking's claim that certain human kinds cannot be natural kinds because they are affected by classificatory feedback effects. Critics have rejected this claim. They argue that certain biological kinds – which, on the account of natural kinds at issue, are paradigmatic natural kinds – participate in the very same sort of feedback effects as human kinds. I object that the extant discussion is flawed in two ways: it relies on an overly simplistic account of natural kinds as stable vectors for projections, and it does not acknowledge that the human kinds in question are hybrid kinds that pose special challenges to scientific understanding. I argue that natural kinds are better understood in terms of the factors and mechanisms that support patterns of change and stability among the members of the kind. Against this background, I identify several reasons why hybrid kinds might make poor candidates for natural kinds thus understood.

Due to their structure of base and associated status, hybrid kinds can simultaneously be the subject of natural scientific and social scientific inquiry. In order to understand how social and natural scientific explanations relate in such a constellation, Chapters 5 and 6 use our previous findings regarding the ontological nature and epistemic challenges of hybrid kinds to advance a longstanding, heated debate at the crossroads of the natural and social sciences: the discussion on biological versus social explanations of gender differences in human behaviour and psychology. Chapter 5 focusses on the issue of causal

explanation. Evolutionary psychologists propose that many psychological or behavioural differences between men and women are the product of evolved psychological mechanisms. Feminist theorists object that this research ignores environmental determinants of gender differences while naturalising and thus reinforcing the oppression of women. I argue that the extant debate fails to pinpoint the causal-explanatory conflict between these two approaches. To move forward, I suggest to distinguish disagreements about causal facts from disagreements about the requirements of good explanation. I argue that disagreements about causal facts occur only when evolutionary psychologists make claims about proximate mechanism. Such claims, I point out, are not usually warranted by evolutionary psychology's methodology, because feminist inspired work in the social sciences suggests that important preconditions are not fulfilled.

Chapter 6 continues this inquiry by looking at disagreements about explanatory pragmatics and its relation to feminists' political concerns. Starting from the observation that many feminists tend to ignore evolutionary psychological explanations of gender differences altogether, I explore the potential for a "right to ignore" the relevant causal claims irrespectively of whether they are true or false. I argue that Maria Kronfeldner's argument for a right to ignore, which centres on Alfred Kroeber's defence of the autonomy of cultural anthropology from biological anthropology, cannot be applied to the feminist case. I consider two further common attempts to justify feminists' ignoring of biological explanations and argue that both of them are problematic. Against this background, I develop an alternative justification. I suggest that feminists are often justified in ignoring biological explanation of gender differences because a whole range of feminist demands are largely independent of questions of natural difference.

1

ARE SOCIAL KINDS SPECIAL?

REFLEXIVITY AND THE HYBRID KIND MODEL

What, if anything, makes social kinds different? Over the course of this thesis, we will encounter a number of different answers to that question. As a starting point, however, it makes sense to consider an account that predates and influenced many recent theories of social kinds: John Searle's account of social kinds. Searle's account made prominent the idea that social kinds are conventional or mind-dependent in a specific way– they are *reflexive*.[1] Call this the *reflexivity account* of social kinds.

Saying that a kind is mind-dependent simply means that the existence of the kind somehow depends on the existence of human minds. This feature is shared by a number of kinds which are not social kinds – such as mental states like pain – and is thus not particularly interesting in this context. Reflexivity, however, is a very specific form of mind-dependence. For Searle, to say that a kind is reflexive, generally speaking, is to say that the existence of the kind somehow depends on human beliefs, propositional attitudes, or intentional states about the kind itself.

Searle's account has made popular the idea that social kinds are epistemically and ontologically different from other kinds. The basic idea is that, ontologically, many social kinds are the way they are because we decided that they be like that. This idea is expressed in the form of metaphors suggesting that these kinds are "linked" or "glued together" by human convention, or that they lack "natural boundaries." I will refer to this idea as the *lack of natural cohesion* of social kinds. From this alleged lack of natural cohesion, several philosophers have concluded that these social kinds must also crucially differ from natural kinds epistemically. Many social kinds, they have suggested, are *epistemically transparent*, meaning that these kinds fail to tell us something about the structure of the world because we cannot be wrong about them.

Others have challenged this view. Francesco Guala argues that human beliefs about social kinds are neither necessary nor sufficient for their existence, and the associated claims about lack of natural cohesion and epistemic transparency are false. Guala suggests that social kinds are mind-dependent only in a minimal, linguistic sense: we cannot be

[1] Searle also refers to these kinds as "self-referential".

wrong about what counts as "money" in a folk conceptual sense that we collectively define, but that tells nothing about the kind *money* that has a central role in the social sciences.

In this chapter, I argue that neither Guala nor his opponents provide a convincing account of the social kinds in question. Although Guala rightly rejects the predominant interpretation of Searle's account, he ignores important insights from Searle's original characterisation of social kinds. I argue that a more promising understanding needs to focus on the fact that these social kinds involve a special sort of relation, by which social statuses or properties are imposed (or conferred) onto existing entities. On the basis of these insights, I develop my own model, which understands the kinds in question as *hybrid kinds*. I then use the hybrid kind model to clarify claims about the ontology and epistemology of the kinds in question. For that purpose, I will first introduce Searle's understanding of social kinds as reflexive kinds and the associated claims about their lack of natural cohesion and epistemic transparency. I then introduce Guala's objections to this view and the alternative view he defends. After pointing out several problems with both accounts, I develop my own account of hybrid kinds and explore how it can be used to clarify the extant discussion.

1.1 SOCIAL KINDS AS REFLEXIVE KINDS

1.1.1 Searle's account of social kinds

The idea that social kinds depend on human beliefs goes back a long way.[2] In more recent history, the most detailed and influential account of this has been formulated by John Searle (Searle 1995, 2010). Searle's (1995) account of social kinds seeks to make sense of the puzzling observation that social kinds like money, property, governments and marriages seem to exist only because humans believe them to exist.[3] The reason for this, according to Searle, is that such kinds are created by the (ongoing) collective acceptance of *constitutive rules* that have the form "X counts as Y in context C" (Searle 1995, 28). Via this process, existing objects are assigned a new status and accompanying functions that the phenomena could not perform in virtue of their pre-existing physical features – an "institutional" kind is created (Searle 1995, 46). For example, in the case of money (or,

[2] Epstein (2015) observes that it goes at least back to David Hume,

[3] Searle refers to these entities as "social facts" and "institutional facts" respectively. Since many of his examples are more appropriately described as kinds, and since this thesis is largely concerned with kinds I, I will reformulate his remarks in terms of kinds where necessary. Searle also refers to social kinds as "institutional kinds".

more precisely, legal tender in the United States), Searle suggests that the relevant constitutive rule is something like the following:

> *Money*: Pieces of paper that are made from a specific material, have a specific physical design, and have been issued by the Bureau of Engraving and Printing (BEP) count as money in the United States.[4]
>
> (Searle 1995, 45-6)

Importantly, according to Searle, the existence of kinds like *money* therefore depends on people having propositional attitudes about money. In order for something to be money, people need to believe that it is money. Searle describes this phenomenon as the "self-referentiality of social concepts" (Searle 1995, 32, 53). In the following, I will refer to the kinds that fit this characterisation as *reflexive kinds*.

Searle suggests that we can distinguish between two sorts of reflexive kinds. For some kinds, such as *money*, the existence of the kind or type depends on the human beliefs that the thing in question is *money,* yet individual tokens or instances of *money* can exist without being recognised as money. Call these sort of kinds *type-reflexive*. According to Searle, a dollar bill which falls straight from the printing press into a floor crack is money even if it is never recognised or used by anyone as money. At the same time, a circulated counterfeit dollar bill fails to be money even if no-one ever recognises it as counterfeit. Searle distinguishes type-reflexive kinds from a second sort of reflexive kinds for which both the existence of the kind (or type) and the existence of each individual instance (or token) depends on the human belief that the thing in question is or is not of the kind in question. As an example, Searle discusses the kind *cocktail party*. He suggests that an individual event is not a cocktail party if no-one thinks of it as a cocktail party. In other words, kinds of this sort are both type- and token-reflexive. For brevity, I will refer to them as *token-reflexive kinds*.

1.1.2 The ontological and epistemological implications of reflexivity

Searle's account of social kinds as reflexive kinds has been highly influential. Although many people have argued, *pace* Searle, that not all social kinds are reflexive kinds, reflexivity is still commonly thought of as a key difference between at least some social kinds on the one hand and natural kinds on the other (Thomasson 2003, Khalidi 2015). Amie Thomasson points out that, while it might be plausible to say that all social kinds are mind-dependent – i.e. all social kinds depend on human mental states of some sort –

[4] This constitutive rule is commonly abbreviated in a way that only refers to the condition of being printed by the BEP, see below.

not all social kinds depend on human intentional states about the kinds (Thomasson 2003, 585n8, 606). For example, kinds like *bigotry*, *racism*, or *recession* depend on certain mental states about other people or financial transactions, but those mental states need not be about the kinds themselves.[5] In other words, Thomasson suggests a two-way distinction of social kinds between reflexive kinds on the one hand and non-reflexive kinds on the other. Muhammad Khalidi, another recent proponent of Searle's view of social kinds, goes one step further than that. Taking up Searle's distinction discussed above, he proposes a three-way distinction that distinguishes non-reflexive, type-reflexive, and token-reflexive kinds (Khalidi 2015).

Of course, these classifications are not taxonomic ends in themselves, but are meant to demarcate important epistemic and ontological differences between different types of social kinds. Epistemically, reflexivity has been argued to entail *transparency*. According to Thomasson (2003), our knowledge about reflexive kinds is immune to certain forms of ignorance or error that (non-reflexive) natural kinds are susceptible to. More precisely, Thomasson argues that certain realist epistemological commitments associated with natural kinds do not fully apply to social kinds which are reflexive.[6] Consider the example of money. According to Thomasson, the existence of money depends on the collective acceptance of the constitutive rule "bills printed by the BEP count as money in the US". However, this rule establishes necessary and sufficient conditions for being money. As a result, she suggests, we can neither be wrong about the conditions for being money, nor can we be ignorant about these conditions given that money exists.

Thomasson points out that her transparency thesis has important limitations. For one thing, it does not apply to non-reflexive social kinds. People can be racist, or a society can be in a recession, without anyone having beliefs about racism or recession. Epistemically, non-reflexive social kinds therefore differ remarkably from reflexive social kinds. While we cannot be ignorant or wrong about the nature and existence of reflexive kinds, knowledge of both the existence and nature of non-reflexive kinds depends on

[5] The examples of racism might be a somewhat misleading example. On most account of racism, for racism to exist it is necessary that people believe (or did at some point believe) that humans come in distinct races and that some of these are superior to others. This is identical with some one-line dictionary definitions of racism. If we take these definitions at face-value, people who belief that humans come in superior and inferior races and are aware of that have a concept of racism (contrary to Thomasson ibid; Khalidi 2015, 100n5). However, the concept of racism is usually (intended to be) a thicker one including the ideas that the belief in racial superiority is based on scientifically wrong assumptions, that it is morally objectionable, that is has specific political and economic functions, etc.. This "thick" concept of racism is obviously not necessary for the existence of racism.

[6] Thomasson makes the same observation for semantic commitments (such as the causal theory of reference). Since the focus here is on epistemic features of social kinds, the discussion is limited to these. For a discussion of the semantics of social kinds, see Chapter 3.

substantive scientific discovery. In addition, Thomasson emphasises that transparency does not imply that it is impossible to acquire any new knowledge about reflexive kinds. Although we cannot be wrong or ignorant about the nature of money, we can still have meaningful social scientific inquiry about the "causal relations" involving money, such as the "(perhaps) unintended and unnoticed oppressive consequences of our practices involving money" (Thomasson 2003, 606). In other words, while the existence of a reflexive social kind implies that we are aware of its ("definitional") nature, empirical inquiry may be necessary to reveal its causal properties.

Khalidi provides a very similar account of the epistemic consequences of reflexivity. He suggests that reflexivity marks the difference between kinds which are suitable objects for scientific inquiry and those which are scientifically interesting only to the extent that they participate in "new causal patterns". However, unlike Thomasson, who thinks the crucial epistemic distinction lies between reflexive and non-reflexive kinds, Khalidi argues that it lies between type-reflexive and token-reflexive kinds. According to Khalidi, non-reflexive and type-reflexive kinds function epistemically like natural kinds because their properties are subject to scientific discovery. Token-reflexive kinds, by contrast, can be subject to scientific discovery only to the extent that they participate in "new causal patterns", i.e. are associated with properties that have not been "written directly into the category itself" (Khalidi 2015, 106-7). As an example of token-reflexivity, Khalidi discussed the kind permanent resident. He argues that it is entirely up to us (or our governments) what properties an individual needs to have to qualify as a permanent resident – we could decide that they need to be capable of swimming hundred metres underwater. Yet permanent residents can enter into novel causal patterns (such as mainly settling in urban areas) that can be discovered by empirical inquiry.

Consider now the ontological implications of reflexivity. Thomasson and Khalidi both argue that the epistemic distinctions they identify reflect important ontological differences between different types of social kinds. According to Thomasson, reflexive kinds lack "natural boundaries." She suggests saying that a kind has "natural boundaries" means that the kind corresponds to a structure in the world that is independent of our beliefs about it. It implies that the kind "is not merely a division artificially imposed on the world by human concepts" (Thomasson 2003, 582). In other words, for Thomasson, both the existence and the boundaries of a reflexive kind depend on human beliefs about the kind.

Khalidi provides a somewhat different story. According to him, the epistemic distinction he identifies is based on a central ontological difference in terms of what

connects or "unifies" the properties that characterise the social kind. While the properties that characterise non-reflexive and type-reflexive kinds are (at least partly) causally connected, the properties associated with token-reflexive kinds are linked by human convention i.e. are associated with the kind because of "social rule or convention" (Khalidi 2015, 106). More precisely, Khalidi suggests that non-reflexive kinds have only causally connected properties, type-reflexive kinds have both causally and conventionally connected properties, and token-reflexive kinds have only conventionally connected properties.[7] Since, according to Khalidi, non-reflexive and token-reflexive kinds are at least partly understood in terms of causal properties, they function epistemically like natural kinds. Token-reflexive kinds are different because the properties associated with these kinds are not related by causal connections but merely by human conventions.

Let's briefly recapitulate, Thomasson and Khalidi both assume that there is a "realist" epistemic characteristic of natural kinds whereby what it means to be of the kind is subject to scientific discovery rather than human stipulation. They also agree that, while some social kinds share this feature, others lack it. These latter kinds, they argue, can be subject to scientific discovery only to the extent that they participate in unintended causal relationships. Finally, they both suggest that this epistemic distinction reflects important differences at the ontological level. The reason some social kinds cannot be subject to scientific discovery is because their "boundaries" or "unity" do not reflect the independent structure of the world but are a matter of human convention.

At the same time, Thomasson and Khalidi disagree about where exactly these crucial epistemic and ontological differences can be found. For Thomasson, the distinction lies between non-reflexive and reflexive kinds; for Khalidi, between type-reflexive and token-reflexive kinds. Furthermore, when discussing the underlying ontological differences, they not only disagree about where these differences ought to be located, but also – as far as the metaphorical language allows to conclude – about what exactly these differences consist in. While Thomasson believes the difference is a matter of the kinds having natural rather than conventional *boundaries,* Khalidi suggests it is a matter of having causal as opposed to conventional *linkage* among the kind's properties. However, I suggest we put these minor differences to one side and instead recognise that both Thomasson and Khalidi are essentially suggesting that the properties that constitute (token-) reflexive kinds are held together by convention. I will refer to this idea as *conventional cohesion*. The

[7] Khalidi refers to the former two as *causal kinds* and to the latter as *conventional kinds*.

concept is a vague one, but it suffices to identify the intuition that underlies both Thomasson's and Khalidi's account.

To summarise the previous discussion, we can identify three core assumptions about the nature of reflexive kinds. These are:

(i) *Conventional Cohesion:* The properties that constitute the kind are "held together" by our beliefs.

(ii) *Epistemic Transparency*: We cannot be wrong or ignorant about the nature of (token-) reflexive kinds.

(iii) *Type-token Reflexivity:* Reflexive kinds fall into two types, some of which are type-reflexive and some of which are token-reflexive.[8]

Not everyone agrees with the reflexivity account of social kinds. More recently, Guala has argued that we should wholeheartedly reject the reflexivity view and replace it with a realist account of social kinds that is informed by game theory (Guala 2010; 2014). As part of this proposal, all three core assumptions about reflexive kinds have come under attack. In the following section, I recapitulate and discuss Guala's objections.

1.2 GUALA'S OBJECTIONS

According to Guala, the reflexivity view of social kinds is misguided because it does not provide a convincing account of its paradigm examples. Those social kinds that have been described as reflexive kinds are not reflexive at all – human propositional attitudes about them are neither necessary nor sufficient for their existence. To show this, he invokes the case of *money*, which has been a paradigm sample of reflexivity since Searle introduced the view. Guala argues that we can see that the belief that something is money is not sufficient for the existence of money by imagining a case of hyperinflated currency. In the context of hyperinflation, people might continue to refer to bills printed by the BEP as "money" even though they have ceased to use it as currency and used cigarettes instead. According to Guala, although they are still called "money", these bills no longer perform the core functions of money as identified by economists: they are no longer used as a means of exchange, store of value, or unit of accounting. As a result, the bills are no longer money in the relevant sense of the term used by social scientists.

[8] Thomasson recognises the distinction between type-reflexivity and token-reflexivity, but argues that it is inconsequential for the epistemic and ontological properties of the kinds (Thomasson 2003, 586).

According to Guala, this shows that people's belief that bills printed by the BEP are money is not sufficient for making the bills members of the kind *money*. The fact that these beliefs are also not necessary, he argues further, becomes obvious when we probe the notion of collective acceptance in more detail. To see this, it is helpful to distinguish between three notions of collective acceptance (Guala 2010, 252-8). The first one is the *full-transparency* notion of collective acceptance. It presupposes that people in a society know not only the conditions for membership in a reflexive kind, but they also know that the existence of the reflexive kind is a matter of their collective acceptance of these conditions. The second is a somewhat weaker *cognitivist* notion of collective acceptance. According to the cognitivist notion, individuals in a society need to be aware of (that is, consciously accept) the conditions for membership in a kind, but they can be ignorant of the constitutive role of the collective acceptance of these conditions. For instance, people need to collectively accept that a bill has to be printed by the BEP in order to be money, but they may be ignorant of the fact that the existence of money depends on their collective acceptance of this condition. Thirdly, there is the *non-cognitivist* version of collective acceptance. This version states that collective acceptance of membership conditions is a matter of implicit rather than conscious acceptance. On the non-cognitivist account, collective acceptance can be expressed in terms of patterns of action, for instance by individuals treating objects that fulfil certain conditions as members of the kind. But it does not require that these individuals have any conscious representation of what the conditions for membership in the kind are.

As Guala points out, although proponents of the reflexivity account of social kinds need not be committed to the full-transparency notion, they need to endorse a cognitivist rather than a non-cognitivist notion of collective acceptance. The reason for this, he argues, is that a non-cognitivist notion of collective acceptance is incompatible with the reflexivity view's core tenet that the existence of social kinds depends on propositional attitudes. Accepting a non-cognitivist notion of collective acceptance entails that collective acceptance can be a matter of patterns of action rather than propositional attitudes about the kind. Hence, if proponents of the reflexivity view were to endorse a non-cognitivist notion of collective acceptance, they would have to admit that propositional attitudes about reflexive kinds are not necessary for the existence of these kinds after all.

More important for our purpose, Guala points out that a non-cognitivist notion of collective acceptance cannot support the epistemic points that proponents of the

reflexivity account are trying to make (Guala 2010, 258). Saying that reflexive kinds can exist without anyone having any propositional attitudes about them is effectively the same as saying that everyone can be ignorant about the existence or nature of these kinds. In other words, non-cognitivist collective acceptance is in direct conflict with the reflexivity view's core assumption of epistemic transparency.

But not only that. In addition to Guala's point about epistemic transparency, we can note that a non-cognitivist notion of collective acceptance would proof fatal to the other two core assumptions about reflexive kinds as well. This is because Conventional Cohesion and Type-token Reflexivity both depend on people having propositional attitudes about kinds or their members. Without anyone holding such attitudes, it can neither be the case that the conditions that constitute a reflexive kind are somehow unified or "held together" by people's beliefs about them (Conventional Cohesion), or nor can there be a distinction between reflexive kinds that require beliefs about tokens and reflexive kinds that only require believes about types. In other words, on a non-cognitivist understanding of collective acceptance, Conventional Cohesion and Epistemic Transparency, and Type-token Reflexivity all fail.

According to Guala, this is bad news for proponents of the reflexivity view: all three core assumption have to be rejected because the non-cognitivist notion of collective acceptance is the only plausible one. This, Guala argues, is because it is a truism that we can be wrong about the membership conditions of reflexive social kinds – we can wrongly believe that money is anchored by a gold standard, that the king is divinely ordained, or that witches have made a pact with Satan (Searle 1995, 47; Guala 2010, 256-9; 2014, 3). These examples of reflexive kinds show that we evidently hold factually false beliefs about such kinds. In some cases, such as for the kind *witch*, the beliefs in question could not possibly be true.

On the basis of these arguments, Guala concludes that the reflexivity view of social kinds ought to be rejected. What does he make, then, of the alleged difference between those social kinds that have been classified as "reflexive" on the one hand, and non-reflexive social kinds and natural kinds on the other? Was the intuitively so compelling idea that money depends on human beliefs in a way that recessions and tigers don't just a mirage? In lieu of the reflexivity account Guala offers a proposal that acknowledges a distinction between kinds that are belief-dependent and kinds that are not. But his proposal spells out the distinction in a very different way. According to Guala those kinds that were previously thought of as reflexive kinds – kinds that are constituted by beliefs

about them – are better understood as being constituted by "systems of beliefs and actions in equilibrium" (Guala 2014, 3). He suggests that

> What matters is not what type of attitude people have toward a certain class of entities (the conditions they think the entities ought to satisfy in order to belong to that class), but what they do with them in the course of social interaction. The relevant attitudes, in other words, are directed toward the attitudes of other people.
>
> (Guala 2014, 5)

In other words, Guala recognises a type of social kinds which are constituted by beliefs. Yet, *pace* the reflexivity account, he thinks these beliefs are not about the kinds themselves, but rather about the beliefs of other people, for example whether they will accept certain paper bills in exchange for goods and services.

The collectively accepted beliefs philosophers have traditionally identified as constitutive of social kinds (such as "bills printed by the BEP are money") play a very different role in Guala's view. They are beliefs about *coordination devices* that facilitate the convergence of actions and beliefs which constitutes a social kind (Guala 2014).[9] To understand this point, consider the example of money. In Guala's view, in order for something to function as money, there needs to be a stable system of beliefs that others will accept certain items in exchange for goods. For such a stable system of beliefs to arise and remain, it helps if individuals agree on a "signal" that allows them to reliably predict each other's beliefs and actions. In the case of money, Guala suggests, being printed by the BEP acts as such a coordination device, signalling that when presented with such a bill, others will accept it in exchange for goods.

There are several problems with Guala's account. For one thing, it relies on a game-theoretic understanding of social life that isn't universally shared (Hargreaves Heap & Varoufakis 1995). Since this debate would take us to far afield, I will not get into any details. A problem with Guala's account that is much more pertinent to the discussion here, however, is that his account of social kinds does not apply to many kinds that we would intuitively want to include among the "belief-dependent" or "reflexive" kinds. It has been argued that racial kinds like *black* can plausibly understood as social positions that are constituted by being oppressed or discriminated against on the basis of skin colour (Haslanger 2012). This suggests a striking parallel between *money* and *race*. Just like *money* can be understood as patterns of actions and beliefs that are coordinated around specific types of paper, *race* can be understood in terms of patterns of beliefs and actions

[9] Note that Guala uses the terms "coordination device" and "correlation device" synonymously.

that are coordinated around individuals that have certain physical features. However, whereas in the case of *money*, the relevant beliefs are about the beliefs and actions of other people (as Guala's account requires), in the case of *race*, the beliefs are typically erroneous beliefs about the kind or tokens of the kind.

To see this, consider in more detail the kinds *money* and *race*. In the case of *money*, we have seen above that the patterns of selling, buying, exchanging, etc., that constitute *money* may involve erroneous beliefs about the kind or its tokens, such as the belief that paper bills are anchored by the gold standard. However, according to Guala's account, these erroneous beliefs do not by themselves produce the patterns of actions and beliefs that constitute money. To collectively create a stable system of buying, selling, and exchanging around these paper bills, people need to believe that other people will accept the bills in exchange for goods and services, now and in the foreseeable future.

This is the main point of Guala's game-theoretic account of social kinds. However, it is not the case for the example of *race*. Here, it is plausible to argue that the relevant patterns of beliefs and actions arise solely on the basis of individuals' erroneous beliefs about the kind or its tokens. While it is important to recognise that racism is not simply a matter of individual prejudice, but a complex interplay of economic, cultural, and legal factors, it is possible for individual racist attitudes alone to give rise to relevant patterns of oppression and discrimination. For instance, a widespread belief that people with dark skin colour are less professionally capable than people with light skin colour can lead to a situation where the former people systematically end up in less well-paid and less prestigious professions than the latter. In this example, *black* then is a social kind constituted by a pattern of professional discrimination, and this social kind is associated via human beliefs with individuals who have dark skin. However, unlike *money*, the pattern that constitutes *black* does not require beliefs about the actions and beliefs of other people.[10] Real-world cases of racism are certainly more complex than this and might be amplified by beliefs about the racist beliefs or actions of other people (e.g. "If I hire someone with dark skin, my racist customers will stay away"). Nevertheless, on a theoretic level, such kinds can be constituted by patterns of beliefs and actions that simply emerge as the aggregate of actions based on individual erroneous beliefs, i.e. their racist prejudices about the capabilities of people with black skin. Analogous arguments, I believe, can be made in the case of *gender* and *witches*.

[10] It might involve racist beliefs about the actions of other people such as "a dark-skinned person will not perform well at this job", but these are not the beliefs that are at issue in Guala's game-theoretic model.

A possible reaction to this would be to accept Guala's account but exclude kinds like race, gender, and witches as a different type of phenomena that require a model of their own. This move, I believe, comes at too high a cost. The parallels between social kinds like *money* on the one hand and kinds like *race* and *gender* on the other are too cunning to be ignored. In both cases, a central feature of the phenomena in question is that certain classifications of entities (metal coins, humans of specific body types) are associated via human conventions with specific social functions or statuses. Accordingly, to formulate a unified account might improve our understanding of examples that fall into either category. This is not to say that Guala's account should be rejected as a useful and illuminating way of understanding some social kinds, or at least some aspects of some social kinds, that have traditionally been understood as reflexive kinds. What I am hoping to show, however, is that when it comes to making sense of the phenomena traditionally referred to as "reflexive" kinds Guala's account isn't the only game in town. To the contrary, his account neglects some striking features of these kinds that deserve a model of their own.

1.3 REFLEXIVE KINDS AS HYBRID KINDS

While Guala's proposal may elucidate one aspect associated with "reflexive" kinds, I will propose an alternative account that has several advantages. It is more general in two respects: it covers cases like *race* and *gender* in addition to those addressed in Guala's proposal and it is not committed to a particular game-theoretic view of social reality. Moreover, it picks up a familiar element of social ontological theorising: its core aspect is expressed in Searle's claim that reflexive kinds involve the imposition of a status function onto an existing object. In the following, I will first describe Searle's discussion of status-endowment and point out the overlap with Asta Sveinsdottir's recent account of conferred properties. I argue that both accounts can be combined into a more general account of reflexive kinds as *hybrid kinds* that involve positioning existing objects or individuals in a new network of social relations. I then show how this account can be used to clarify the extant discussion on reflexive kinds, while avoiding the unnecessarily restrictive commitments of Guala's account.

1.3.1 Reflexivity's lost brother – the idea of imposed status functions

To develop this alternative account, we first need to distinguish two aspects in Searle's (1995, 2010) characterisation of reflexive kinds outlined above. On the one hand, there is the familiar idea that reflexive kinds depend on human beliefs about them. As evidenced

in the previous discussion, this idea has attracted the majority of philosophical attention. On the other hand, there is the less noted idea that such kinds involve the imposition of a status function onto an existing (kind of) object, individual, or event. So far, we have been focussing on the aspect of reflexivity because this aspect dominates the discussion on Searle's account of social kinds and is considered as its defining characteristic (hence the terminology of "reflexive" kinds).

In some respects, this is not surprising. The realist paradigm in philosophy traditionally understands its subject matter in terms of what exists independently of human belief, so as to distinguish the "real" from the "fictional". Against this background, the observation that certain social kinds seem to depend on human beliefs about them, yet look so much unlike dragons and unicorns, can seem highly peculiar. If you think what is most striking about social kinds like money and requires explanation is their perceived mind-dependence, it makes sense to define social kinds along that dimension. Most scientific and philosophical inquiries start with such "explorative" definitions of the target phenomenon. But explorative definitions typically leave a lot of room for improvement in the light of new insights. In the case of reflexive kinds, for example, we would want to know things like "Why are reflexive kinds dependent on beliefs about them?" or "What about reflexive kinds is dependent on such beliefs?" Most importantly, we would want to know how exactly the idea that these kinds are dependent on human beliefs relates to Searle's other idea that these kinds involve imposed status functions. What is surprising in the case of reflexive kinds, then, is not that they have originally been characterised by an explorative definition focussing on their alleged reflexivity. What should surprise us is that this explorative definition has developed a life of its own as the dominant understanding of the phenomenon in question, while the idea of imposed status functions, and the question whether and how the idea of imposed status functions can be squared with the dominant understanding, has been largely disregarded. This, I will argue, has been much to the detriment of the extant discussion. We can get a much firmer understanding of the phenomenon in question if we explore the idea of imposed status functions instead.

Let's start by considering in more detail what Searle has to say on the matter. Searle, as argued above, believes that a reflexive kind is created by the collective acceptance of a constitutive rule of the form "X counts as Y in C".[11] What exactly is going on here? Searle

[11] I abbreviate this enumeration in the following by referring only to objects. Individuals and events are implicated by that, except where the context makes is clear that this is not the case.

suggests that what the constitutive rule describes is the imposition of a so called *status function* onto existing objects, individuals, or events. In the formula, X refers to (a type of) existing objects, Y refers to the status function, and C to a specific context in which the constitutive rule applies. To clarify the idea of a status function, Searle distinguishes it from what he calls *causal* functions. According to Searle, status functions differ from causal functions in terms of what needs to be the case so that an object can perform the function. Objects can perform a causal function simply in virtue of their "intrinsic physical features" (Searle 1995, 39). Status functions, by contrast, cannot simply be performed in virtue of an objects existing features. Instead, they require our help in the form of collective acceptance or intentionality. A status function is a function that can only be performed because people collectively assign a certain status to that object.

Searle illustrates the difference with the example of a stone wall that continues to function as a territorial boundary despite decaying over time (Searle 1995, 39-40). According to Searle, the intact wall can perform the function of a territorial boundary simply in virtue of its physical properties – it physically hinders people from leaving or entering the territory. In other words, being a territorial boundary is a causal function of the wall. The same is not true once the stone wall has decayed to a line of stones on the ground. The wall (or the remains thereof) no longer have the physical properties that stop people from crossing. Yet we would not be surprised to find that people living on both sides of the wall continue to recognise the remains as a territorial boundary, for instance by crossing it only with the allowance of people living within the territory. According to Searle, this illustrates how the function of territorial boundary can be realised in two different ways. Although the line of stones has lost the physical properties that allowed it to function as a territorial boundary (its causal function), it now performs the same role as a status function, i.e. in virtue of collective intentionality that assigns to the line of stones the symbolic status of a territorial boundary.

1.3.2 Asta's conferred property kinds

Asta recently put forward an account that is very similar to Searle's idea of imposed status functions (Asta 2008, 2011, 2013). Although Asta seems to have developed this account independently of Searle's, it provides useful resources for clarifying some of the central concepts and mechanisms we have encountered so far. According to Asta, some social kinds are constituted by properties which are "conferred" onto existing objects, individuals, or events. I will refer to these kinds as "conferred property kinds" in the following. Asta's account of conferred properties has three central elements. The *conferred*

property, which we will specify below, the *grounding property*, i.e. the property that the conferral is attempting to track[12], and a *conferring subject*, i.e. the individual, group, or entity that does the conferring. Her examples of conferred property kinds range from (baseball) *strike* to *woman* and *man*. In the case of strikes, the property of being a strike (conferred property) is conferred onto a pitch that has travelled a certain distance (grounding property) by the umpire (conferring subject). In the case of gender, being a man or a woman (conferred property) is conferred onto people with certain physiological or psychological characteristics (grounding property) by people in their social environment (conferring subjects).

How does Asta's proposal relate to Searle's? Recall that, for Searle, status functions are imposed through collective acceptance of a constitutive rule "X counts as Y in C". At first glance, Asta's concept of a grounding property looks a lot like the X term, her concept of a conferred property looks like Searle's status functions (expressed in the Y term), and the conferring subjects looks like it might have something to do with Searle's condition C. We can confirm this impression by taking a closer look.

Firstly, consider what it is that Asta and Searle claim is conferred or imposed. Since we are already familiar with Searle's idea of a status function, we need to only look at Asta's proposal. Asta's terminology refers to "being a man/woman/strike" as conferred *properties*. This is semantically correct but slightly misleading in that it might suggest that the thing conferred is always a singular or basic property that cannot be further analysed (think of the property "being red"). Asta's discussion makes clear that this is not the case. She argues that conferred properties like *being a woman* are in fact social *statuses* consisting of a set of constraints and "enablements" on an individual's behaviour. In the case of gender, for example, Asta suggests that being classified as a man or woman means to occupy a social status that is characterised by specific duties, privileges and burdens (Asta 2011, 60).

In other words, Searle and Asta agree that, in the social world, there exists a special phenomenon by which social statuses are imposed or conferred onto existing objects, events, or individuals. What exactly then is a social status? It seems that Searle and Asta are disagreeing. Searle, as argued above, suggests that statuses are associated with a number of functions. The status money, for instance, is associated with the function of being a means of exchange, a storage of value, etc. (Searle 1995, 46). By contrast, Asta

[12] Note that Asta's terminology of "grounding properties" ought not to be confused with the notion of metaphysical grounding, which we will encounter in Chapter 2. The properties Asta is referring to have no metaphysical grounding function in the context of her account.

suggests that statuses should be understood in terms of constraints and enablements. I propose that we need not decide between these two proposals, but can accommodate them into a more general account.

1.3.3 Hybrid kinds

According to my proposed account, a status consists of a set (or cluster) of social properties which are acquired by the process that Searle and Asta have described as "conferring" or "social imposing." I take it that the notions of conferring and imposing are largely synonymous in this context, and will accordingly use them interchangeable in the following. The central idea seems to be that, by conferring a certain status on an existing object, people endow the object with new social properties that it previously lacked. To endow an object with a new social property means, essentially, that people relate to the object in a different way. Their attitudes towards the object change, either giving rise to new patterns of behaviours or coordinating and stabilising existing patterns of behaviour involving the relevant object. Social properties, on my proposed account, include functional properties on the social level (e.g. being a means of exchange, being a mentor), social constraints, privileges and expectations (e.g. not being allowed to be aggressive, being able to access child care, being expected to be a criminal), and possibly others. They are properties which are, at some level, constituted by social practices and can therefore only be exemplified in the context of social organisation. This idea, I believe, is captured by Searle's suggestion that our ideas of these statuses are just "placeholders for patterns of activities" (Searle 1995, 57). Entities that share a specific status, i.e. a specific set of conferred social properties, can be grouped into a kind, and I will refer to these kinds as *status kind*s.

Now that we have a firmer idea of what conferred statuses are, we can consider Asta's and Searle's proposals in more detail. Asta proposes that we should distinguish between two types of conferral. Statuses can either be conferred by some person or entity who is authorised to make that conferral, or they can be conferred by the attitudes and actions of people without drawing on any special authority. Asta refers to these processes as *institutional* and *communal* conferring (Asta 2017). Communal conferral occurs in the case of gender, for instance when people collectively identify a certain individual at a party as a woman. The property *being a strike* is an example of institutional conferral. According to Asta, *being a strike* is not conferred by collective acceptance, but is first of all an act carried out by the umpire, who has been authorised to make these conferrals.

Again, it might seem that there is a disagreement here: while Asta argues that we need to distinguish between institutional and communal conferring, Searle seems to suggest that conferring is always a matter of collective intentionality. But this disagreement is only superficial. In fact, both Asta and Searle acknowledge that there can be a cascade of conferred properties. Searle suggests that the Department of Motor Vehicles (DMV), which can (institutionally) confer onto individuals the status *licensed driver*, is itself constituted by a conferred status. Society has (communally) conferred onto the individuals who form the DMV the property *issuing driver's licenses* (Searle 1995, 106). This is analogous with Asta's baseball example, where the property *being an umpire* is also an institutionally conferred property, which has been conferred onto an individual by, say, the supervisory staff of Major League Baseball, and so forth. Although the details of these cases are debatable (is *issuing a driver's license* really communally conferred onto the DMV, or is it institutionally conferred by an elected government?), we can see that conferring is not always directly a matter of collective intentionality, but can occur in a nested structure. At the same time, since we will eventually run out of authorised people or entities, this nested structure has to peter out in collective intentionality/communal conferral.

So far, I used Searle's and Asta's account of social kinds to clarify the notion of status kind. To further comprehend the phenomenon underlying the discussion about reflexivity and to complete my account of hybrid kinds, we need to distinguish two further central elements. The first element is constituted by the objects onto which statuses are conferred. I will refer to the groups or types of objects that have a specific status conferred onto them as the *base kinds*. A base kind is constituted by objects, events, or individuals that have specific *base properties*. Following Asta's idea of grounding properties, I assume that base properties are those properties that the conferring subjects try to track when they confer a social status.[13] We have already encountered several examples of base kinds in the previous discussion. In the case of *money*, Searle's account suggests that the base kind consists of bills printed by the BEP. Asta argues that human individuals having a specific role in biological production constitute the base kind for *men* and *women*, and so forth.

Secondly, and more interestingly, there is the relation connecting the base kind and the relevant social status. Both Searle and Asta notice that there is something peculiar about

[13] I deviate from Asta's terminology of "grounding properties" so as to avoid confusion with the metaphysical notion of grounding (see Chapter 2).

this relation. Searle describes it as a "nonphysical", "noncausal" and "incidental" (Searle 1995, 42, 49). Similarly, Asta suggests that

> there is not a fact of the matter as to whether the pitch is a strike or not independent of the judgment of the umpire, but rather it is the umpire's judgment as to the trajectory of the ball that confers on the pitch the property of being a strike
>
> (Asta 2013, 720)

These characterisations hint at something interesting, but they need further spelling out. I suggest one way of appreciating the peculiar nature of the connection between base kind and status kind starts with the observation that having the right base properties is neither necessary nor sufficient for having the relevant conferred status. Having the relevant base properties is not sufficient because without the social act of conferral, the base object does not exemplify the conferred status. It is not necessary either, because objects can acquire the conferred status even when they lack the base properties. As long as subjects do confer a status on an object it does not matter whether the motivating belief that the object has the appropriate base properties is correct or mistaken.

This makes perfect sense in light of what we know now about conferred statuses. As argued above, to have a conferred status essentially means to be subject to specific attitudes and behaviours on a social scale. For instance, what it means for a piece of paper to have the status of money is that people use the piece of paper in exchange for goods, to pay off debts, etc. The piece of paper is money if and only if it has these social properties. In other words, there is nothing other than people's conventions or agreements that connects the property "being a piece of paper printed by the BEP" with the social properties, such as being a means of exchange, that characterise the status *money*. For that reason, I refer to the connection holding between base kind and conferred status as *conventional linkage*.

Putting these ideas together, we obtain an account of hybrid kinds that consists of three core elements: (i) a *base kind*, (ii) a *status kind*, and (iii) a relation of *conventional linkage* connecting the base kind and the status kind. In the penultimate section of this chapter, I use this account to clarify the extant discussion about reflexive kinds.

1.4 RE-EVALUATING THE CORE ASSUMPTIONS ABOUT REFLEXIVE KINDS

Having argued that the notion of hybrid kinds provides a better way of thinking about the phenomena that motivated Searle's account, we can now evaluate the three core claims

that have been made about reflexive kinds in the extant discussion. Recall that these claims are:

(i) *Conventional Cohesion:* The properties that constitute the kind are "held together" by our beliefs.
(ii) *Epistemic Transparency*: We cannot be wrong or ignorant about the nature of (token-) reflexive kinds.
(iii) *Type-token Reflexivity:* Reflexive kinds fall in two types, some of which are type-reflexive and some of which are token-reflexive.

I will first consider Type-token Reflexivity before moving on to Conventional Cohesion and Epistemic Transparency.

1.4.1 Type-token Reflexivity

Type-token Reflexivity, as argued above, states that we can distinguish type-reflexive and token-reflexive kinds. Recall that, for both, the existence of the kinds in question depends on humans having propositional attitudes about the kinds. Yet, it is argued, type-and token-reflexive kinds differ with regard to the existence conditions of their individual members. For type-reflexive kinds, individual members of a kind can exist without anyone believing that they are members of the kind. For token-reflexive kinds this is not the case – their members do not exist (i.e. they do not have instances) unless we think of them as members of a specific kind.

One problem with Type-token Reflexivity that has been ignored in the discussion so far is that it does not sit easy with a fairly uncontroversial assumption about kind ontology. The assumption states that whether or not an object is a member of a specific kind depends on whether or not it has the properties that characterise the kind. According to this assumption, no matter what sort of kinds *money* or *woman* are, something is *money* if it has money-properties, and someone is a *woman* if that person has woman-properties. For type-reflexive kinds, however, the existence conditions for instances of the kind are said to be very different from the existence conditions of the kind itself. It is assumed that instances of the kind can exist without anyone having propositional attitudes about these instances while the kind cannot exist without people having propositional attitudes about the kind. How can this be?

The hybrid kind model allows us to get out of this muddle. The model proposes that the ontological structure of the kinds in question is more complicated than proponents of the reflexivity view have recognised. It suggests that kinds like money are better

understood as hybrid kinds consisting of a base kind that is conventionally linked to a status kind. As a result of this structure, however, the terms that are used to refer to the kinds in question are highly ambiguous. Prima facie, it is not clear whether they refer to the base kind, to the status kind, or maybe even to the whole lot.[14] Depending on how we use the term "money", for instance, we could be referring to the base kind, the status kind, or the whole hybrid kind. The crucial point is that each of these candidate meanings refers to a kind with its own definition or membership conditions. The base kind of money is constituted by bills printed by the BEP, the status kind is constituted by objects that function as a means of exchange, and the whole hybrid kind is constituted by bills printed by the BEP, that, in virtue of conventional linkage, have acquired the social property of being a means of exchange. For each kind in question, however, the uncontroversial assumption holds true. To be a member of the base kind of *money* is to be a bill printed by the BEP, to be a member of the status kind of *money* is to be a means of exchange, and so forth.

This realisation is crucial, because it clarifies how we should think about the thought experiment that motivated Searle's distinction between type- and token-reflexivity in the first place: the bill that fell straight from the press into a crack in the floor. By saying that the bill is money even if it is never used as a means of exchange, proponents of Type-token Reflexivity seem to suggest something like this: by accepting the constitutive rule "bills printed by the BEP count as money", we can confer a status (being a means of exchange) onto a certain type of base kind (bills printed by the BEP) in a wholesale manner. In order for something bills printed by the BEP to have the status of money, we do not need to have certain attitudes towards each individual bill or use every bill for specific actions. My hybrid kind model suggests that this is wrong. Individual objects acquire membership in a status kind in virtue of the fact that they are subject to specific attitudes and behaviours. On this account, the bill that fell into the floor crack is not a member of the status kind *money* because it lacks the defining property of being used as a means of exchange.

Proponents of Type-token Reflexivity could object that the bill which fell into the floor crack nevertheless has the status of money because we have the *disposition* to use it as a means of exchange if we get hold of it. I am inclined to agree with them. Since the patterns of social activities that constitute a status kind might be intermitted rather than ongoing, it seems plausible to allow that the disposition to be subject to certain attitudes or

[14] For a more detailed discussion of the semantics of hybrid kinds, see Chapter 4.

behaviour can be sufficient for being a member of a certain status kind. Of course, the details of such a dispositional account need further clarification, but this would take us too far afield. For now, it suffices to note that reference to dispositional properties does not yield the results that proponents of Type-token Reflexivity are hoping for. The reason for this is because it shifts the defining characteristics of the entire status kind *money* from "being used as a means of exchange" to "having the disposition to being used as a means of exchange". On a dispositional account, the status kind *money* depends for its existence on dispositional attitudes and actions, but so do its individual instances. In other words, Type-token Reflexivity cannot be salvaged by reference to dispositional properties. In lack of a better argument, Type-token Reflexivity should be rejected.

1.4.2 Conventional Cohesion and Epistemic Transparency

Next, consider Conventional Cohesion, which states that the conditions or properties that constitute the kind are "held together" by our beliefs. Drawing on the above observation about hybrid kind terminology, we can now appreciate that this claim is highly ambiguous. Within a single hybrid kind, it is possible to distinguish different sets of properties that each play a different role. On the one hand, there are the properties of the base kind. On the other hand, we have the conferred social properties of the status kind. Taken together, this gives us not one but *three* possible interpretations of the claim that the properties that constitute the kind are "held together" by our beliefs:

(i) *Conventional Cohesion of the base kind:* Our beliefs "hold together" the properties that constitute the base kind.

(ii) *Conventional Cohesion of the status kind:* Our beliefs "hold together" the properties that constitute the status kind.

(iii) *Conventional Cohesion of the hybrid kind:* Our beliefs "hold together" the properties that constitute the base kind with the properties that constitute the status kind.

Before moving on to discuss each of these cases individually, it is worth realising that the problem at hand extends to *Epistemic Transparency*. Epistemic Transparency is the assumption that we cannot be wrong or ignorant about the nature of a reflexive kind. Once we understand reflexive kinds as hybrid kinds, it is not clear whether talk of the "nature" of such a kind refers to the base kind, the status kind, or the entire hybrid kind. As a result, we are again confronted with three possible interpretations of the assumption:

(i) *Epistemic Transparency of the base kind*: We cannot be wrong or ignorant about the nature of the base kind.

(ii) *Epistemic Transparency of the status kind*: We cannot be wrong or ignorant about the nature of the status kind.

(iii) *Epistemic Transparency about the hybrid kind*: We cannot be wrong or ignorant about the nature of the entire hybrid kind.

In other words, to clarify the discussion around the remaining two core assumptions about reflexive kinds, we need to answer two questions: which of the interpretations of Conventional Cohesion and Epistemic Transparency do proponents of the reflexivity view have in mind? And which of these interpretations, if any, is making sensible claims about social reality? As to be expected, there is no direct answer to the first question, because proponents of the reflexivity view do not disambiguate of their core ontological and epistemological claims in the way the hybrid model requires. There is, however, an indirect way to answer this question. We just need to ask which interpretation of Conventional Cohesion and Epistemic Transparency proponents of the reflexivity view would have to commit to in order to support the implications for social scientific practice that they are defending.

As discussed above, Thomasson and Khalidi both suggest that reflexive kinds can be subject to scientific inquiry only in a limited sense: it makes sense to investigate the novel causal relationships that these kinds enter into, but not the "nature" or "conditions for membership" of the kind. Furthermore, when discussing Epistemic Transparency, Thomasson and Khalidi are both explicitly comparing the epistemic role of reflexive kinds with that of natural kinds, which are commonly individuated by their central role in scientific inquiry. This is a role that, in the context of social science, only status kinds fulfil. When social scientists investigate, for instance, the "nature" of money, they generally talk about the status kind, such as the function of money as a means of exchange, rather than specific physical or etiological features of the base kinds.[15] Similarly, social scientists commonly distinguish gender from sex to emphasise that they are concerned with the specific social role that individuals classified as men or women occupy, rather than with the biological differences between the bodies of the people so classified. In the social sciences, the status kinds associated with *money* and *gender* are of central theoretical

[15] Anthropologists might be interested in what sort of objects have been used as money by an ancient civilisation. But when they investigate that question, they are not investigating the "nature" of money but rather the specific details of an ancient economy.

importance. The physical and etiological features of dollar bills or female bodies tend to be secondary.

In other words, Thomasson's and Khalidi's remarks about the epistemic role of reflexive kinds strongly suggest that they are committed to Epistemic Transparency about status kinds. From this, we can infer which version of Conventional Cohesion proponents of the reflexivity view are committed to. We are able to make this inference because proponents of the reflexivity propose a specific logical relationship between Epistemic Transparency and Conventional Cohesion. As argued above, they believe that we cannot be wrong or ignorant about the "nature" of a reflexive kind because the existence of the kinds presupposed that we collectively accept a constitutive rule which sets out the necessary and sufficient conditions for membership in the kind. But this logical inference only works if Conventional Cohesion and Epistemic Transparency refer to exactly the same kind. Therefore, we can conclude that proponents of the reflexivity view must be committed to Epistemic Transparency and Conventional Cohesion about status kinds. Having established which versions of Conventional Cohesion and Epistemic Transparency are endorsed by proponents of the reflexivity view, the next thing we need to do is ask which versions of these assumptions provide an adequate picture of social ontology. I will address the different versions in turn, starting with those endorsed by proponents of the reflexivity account.

1.4.3 Conventional Cohesion and Epistemic Transparency about status kinds

To begin with, consider Conventional Cohesion about status kinds. As it happens, this assumption has already been vehemently criticized in the relevant literature. Rebecca Mason claims that it is not the case that kinds like *money* are constituted by properties that are "gerrymandered", "stipulated arbitrarily" or "at our discretion" (Mason 2016). She argues that we could not simply "stipulate that money is defined by the properties of being blue before time, t, and being two miles from Lake Michigan" (Mason 2016, 842). Instead, what it means to be money is simply to fulfil the three core functions of being a medium of exchange, a store of value, and a unit of accounting.

In a similar vein, Guala points out that economists have long realised that money's core function of being a medium of exchange *causally depends* on its being a store of value. The reason for that, he suggests, is that trade takes place over time. Selling something in exchange for an asset in order to use that asset to purchase something at a later point does not make sense if that asset loses its value in the meantime. Hence, Guala argues, the core functions of money are linked causally and hold "a posteriori, in virtue of the way the

world is" rather than as a matter of arbitrary stipulation (Guala 2010, 260). In other words, both Manson and Guala believe that it is wrong to say that the properties that constitute *money* are in any way held together by us or our attitudes about the kind. Instead, they are determined by the way the social world happens to be.

While these observations may be correct in the case of *money*, they do not seem to apply to all status kinds. Consider Asta's example of a *baseball strike*, a corporate title like *chief executive officer*, or a legal kind like *murder*. In these cases, it seems plausible to say that the properties that constitute the conferred statuses are indeed a matter of human stipulation. The fact that a strike is a move that contributes to a strikeout, by which the batter's turn is over and the other team takes its turn at bat, is explicitly defined in the rules for baseball. The rights, duties, and responsibilities that constitute the role CEO, and how these relate to the structure of a company, are defined by the company's management and usually explicitly written into the company's charter and employment contracts. Similarly, the fact that murder is a serious offence that carries a mandatory life sentence is defined in UK criminal law.

In other words, for status kinds like *strike*, *CEO*, and *murder*, Conventional Cohesion hints at an adequate picture. The picture at hand is what I would call the "social artefact picture" of social kinds: a society collectively decides (or authorises someone to decide) that a certain set of social properties is called "strike", "CEO" or "murder" and then uses these terms to coordinate social life so as to implement or maintain the relevant practices. The properties that constitute the conferred status are stipulated in accordance with specific goals and purposes. Only afterwards is social life arranged so as to implement the properties that were agreed on beforehand. Note that this is compatible with Thomasson's and Khalidi's suggestion that such kinds can participate in novel causal relationships. For instance, we could empirically discover that CEOs are more likely than other groups of the population to have psychopathic traits. Nevertheless, as with artefacts, the fact that the defining properties of CEOs are coinstantiated in the world is a matter of deliberate design.

To sum up, some status kinds are what I have called "social artefact kinds" and for them, Conventional Cohesion does in fact hold true. The properties that constitute the status *strike* or *CEO* or *murder* are, in a sense, "held together" by our beliefs, i.e. our deliberate efforts to organise social life in accordance with a given template, and thereby make it the case that these properties are coexemplified in the required way. Moreover, the logical inference proponents of the reflexivity view make from Conventional

Cohesion to Epistemic Transparency holds for these kinds. Since we have explicitly stipulated the properties of that template, we cannot be wrong about what the properties of the resulting kinds are either.

While the social artefact picture might be adequate for the cases discussed above, it does not apply to all status kinds, let alone paradigm cases in the discussion, such as *money*. For status kinds like *money*, to borrow Searle's terminology, it seems more plausible that the social practices that constitute these kinds simply have "evolved" in the course of human history without any deliberate stipulation or design (Searle 1995, 47). We should also note that there might be borderline cases. Khalidi's discussion of the kind *metic* is insightful here. Khalidi suggests that *metic*, a form of permanent residency in ancient Athens –in other words, a social status kind characterised by rights and duties explicitly stated in Athenian law – might be a formalised versions of social patterns that preceded the existence of the legal kind. According to Khalidi, it is likely that *metic* was modelled on an informal social role which was likely in place prior to the category *metic*. This informal status (call it *proto-metic*) included features like "participation in economic transactions and in military service and non-participation in the political process and in owning property" that were conferred on the basis of prejudice or informal convention (Khalidi 2015, 107). Hence, there might be cases of status kinds that start out as informal patterns of social life and only later become institutionalised and thus a matter of "social engineering". Conversely, there may be examples of social artefact kinds that lose their deliberate enforcement mechanism at some point, yet linger on as informal patterns in social life.

In summary, the discussion above suggests a somewhat limited, accidental victory for proponents of the reflexivity view. There are indeed status kinds for which the assumptions of Conventional Cohesion and Epistemic Transparency hold true. But they have only little overlap with the range of phenomena that the reflexivity account set out to explain in the first place. While Conventional Cohesion and Epistemic Transparency might apply to a subset of status kinds, they are not features of status kinds in general.

1.4.4 Conventional Cohesion and Epistemic Transparency about base kinds

Next, consider Conventional Cohesion and Epistemic Transparency about base kinds. Conventional Cohesion about base kinds states that the properties which constitute the base kind are "held together" by our beliefs. As with status kinds, this claim is true to a limited and somewhat trivial extent because some base kinds may be artefacts that we deliberately design and produce. If we use dollar bills as *money*, the pieces of paper onto which the status of money is conferred are designed and manufactured by us. The reason

that members of the base kind are rectangular slips of paper with specific print on them is because we deliberately made them that way. However, if we would decide to use, say, blue seashells as money instead, this would not be the case. The reason that these objects are blue, shell-shaped and shiny has nothing to do with our attitudes or actions. Again, there might be borderline cases. For instance, we could mark the seashells with a certain pattern before using them, thus making them more artefact like. At the same time, the status functions that we confer on base kind objects put practical limits on our "design" – as Khalidi observed, we could not use ice as money (Khalidi 2015, 105).

In either case, the epistemic implications are modest. We can be wrong or ignorant about the nature of base kinds, even those that are the product of our design. It would be perfectly possible, for instance, to design, produce, and use certain paper bills as money without knowing that the bills are flammable, or that the ink on them is toxic. As a result, both Conventional Cohesion and Epistemic Transparency fail as general features of conferred status kinds.

1.4.5 Conventional Cohesion and Epistemic Transparency about hybrid kinds

The last candidates to consider are Conventional Cohesion and Epistemic Transparency about hybrid kinds, or more precisely, about the relationship between base kinds and status kinds. First, consider Conventional Cohesion about the relationship between base kinds and status kinds. On this interpretation, Conventional Cohesion states what is somehow "held together" by our beliefs is the connection between a base kind and a status kind. Searle seems to take note of this conventional relationship between base kind and status kind. He argues, for instance, that since "the conditions laid down by the X term are only incidentally related to the function specified by the Y term, the selection of the X term is more or less arbitrary" (Searle 1995, 49). (Recall that Searle's X term and Y term roughly correspond to our base kind and status kind.) Unfortunately, neither he nor later proponents of the reflexivity view recognise that this is the only place in their theory where talk about kinds being "held together" by human beliefs makes the most sense.

According to the hybrid model, the properties that constitute the base kinds and the properties that constitute the status kinds are in fact "held together" by our beliefs. The model states that to impose a status onto an existing object is to give that object a position in a network of social relations that it did not previously occupy. But in order for that to happen, people need to recognise or think about the object in a way that triggers a specific range of thoughts or actions in them. Otherwise – and in this sense I would endorse Guala's terminology of "coordination devices" – it would not be possible for people's

object-related thoughts and actions to be brought into line in a way that produces the social patterns that constitute a conferred status. In other words, the connection between base kind and status kind is "held together" by our thoughts because objects of a base kind function as members of a status kind only in virtue of the fact that we think about these objects in a certain way.

What are the implications of this observation? If there is a plausible interpretation of Conventional Cohesion, might the same be true for Epistemic Transparency? Epistemic Transparency about hybrid kinds would suggest that we cannot be wrong or ignorant about the nature of hybrid kinds. The previous discussion suggests that this claim is false. If we can be wrong or ignorant about the nature of base kinds and about the nature of status kinds, it clearly has to be possible that we are wrong or ignorant about hybrid kinds, too. After all, hybrid kinds are nothing other than a compound of base kind and status kind, associated by relation of status conferral that is usually covert and hence something that we are even less likely to understand.

But this is not to say, as Guala seems to suggest, that the kinds in question are just like natural kinds. To the contrary, the fact that the connection between base kind and status kind obtains in virtue of human beliefs has peculiar epistemic consequences that set hybrid kinds apart from natural kinds. Because the social practices that constitute the conferred status do not emerge spontaneously, but require the widespread recognition of, and resulting synchronised or coordinated modification of beliefs and actions towards members of the base kind, recognition *as of a certain kind* is necessary for obtaining the conferred status. And insofar as widespread recognition is automatically accompanied by the emergence of the relevant patterns of social action, it is also sufficient for obtaining the conferred status.

To see this, consider in more detail the sort of beliefs that connect a status kind to a specific base kind. People do not simply believe, say "this is a bill printed by the BEP" or "this is a human with typical characteristics of female anatomy". These beliefs alone would be insufficient to trigger the patterns of thought and behaviour which make a bill function as a means of exchange, or make an individual with female anatomy occupy the social position of women in a specific society. It is only when people recognise a bill printed by the BEP as something above and beyond a bill printed by the BEP, or an individual of female anatomy as something above and beyond an individual of female anatomy, that the relevant patterns of thought and action emerge.

What is going on here, I suggest, is that people need to classify the object of the base kind as a member of a different kind which, in their minds, is associated with a specific concept, i.e. a set of propositional attitudes about the kind. In other words, people classify a paper bill as *money*, or an individual as a *woman* and, by doing that, think about the bill or individuals in terms of the associated concept. The concepts are public in the sense that there tends to be a significant overlap in the propositional attitudes that different people associate with the kinds. Moreover, as evidenced in the previous discussion, the content of the concepts does not have to be true. Guala's discussion of *kings* and *witches* suggests that false beliefs can be just as effective in producing the relevant patterns of thought and behaviour as true ones. In fact, given the complicated hybrid structure of the kinds in question, there is reason to think that they will be false in most cases.

The epistemic upshot of the previous discussion is this: We can be dead wrong about the "nature" of a hybrid kinds like *money* or *woman* in the sense that we can be wrong both about its base properties and its conferred properties. However, in a limited sense, it may be impossible for us to collectively be wrong about the hybrid kind's extension. As argued above, identifying members of a base kind as, say, *money*, *woman*, or *witch,* and interpreting them through the associated concepts, are necessary for creating the relevant status kinds. Accordingly, if enough people come to identify somebody or something as money, a woman, or a witch, and think about and relate to them in accordance with the associated concepts, the individuals/objects will acquire the social status of *money*, *woman* or *witch*. This could make hybrid kinds epistemically transparent in a very limited yet interesting sense: it may be difficult to be collectively wrong or ignorant about which objects are a member of a hybrid kind because what makes these objects a member of the hybrid kind is a matter of people thinking about them in terms of an associated concept. Note that this form of epistemic transparency would be compatible with a limited amount of dispute over kind membership. Having the relevant status does require wide but not unanimous recognition as a member of the kind - we would continue to recognise five pound notes as money even if some shops refused to accept them as a means of exchange.

1.5 CONCLUSION

In this chapter, I argued that the reflexivity view, which understands certain social kinds in terms of belief-dependence, should be replaced with an account of hybrid kinds, which understands them as base kinds associated with a social status. I defended the hybrid model against Guala's alternative proposal by showing that my account is both broader

in scope and does not require us to commit to a specific game-theoretic view of social institutions. Finally, I showed how the hybrid model provides a more nuanced critique of the extant discussion, allowing us to distinguish different interpretations of the core assumptions of the reflexivity view. I argued that Type-token Reflexivity ought to be rejected and that Conventional Cohesion and Epistemic Transparency are not generally true when applied to either base kinds or status kinds. However, I also showed that there is an ounce of truth in Conventional Cohesion and Epistemic Transparency when applied to the relationship between base kinds and status kinds. Hybrid kinds are belief-dependent in the sense that members of the base kind possess the associated status only in virtue of the fact that humans have certain (true or false) beliefs about them. As a result of this, hybrid kinds may be epistemically transparent in the sense that we can be wrong about their nature but we cannot collectively be wrong about their extension.

This shows that the underlying idea that the kinds in question are somehow different from natural kinds is correct. We will return to this idea, and its implications for scientific inquiry, in Chapter 4. Before that, it is time to take a more detailed look at what the hybrid kind model can tell us about the metaphysics of social kinds.

2

GROUNDS, ANCHORS, AND HYBRIDS

THE ONTOLOGY OF SOCIAL KINDS

In the previous chapter, I argued that a common way of thinking about a certain type of social kinds, the reflexivity view, is inadequate. Instead, the kinds in question are better understood as hybrid kinds, which are constituted by a certain status (or status kind) being imposed onto existing objects, events, or individuals (the base kind). In this chapter, I apply the hybrid kind model to a different and more recent discussion in social ontology: the discussion of Brian Epstein's grounding-anchoring model of social ontology.

Epstein proposes a model of social ontology according to which social kinds involve two distinct metaphysical relations: the relation of grounding, which is a familiar concept in metaphysics, and the relation of anchoring, which is a new concept introduced by Epstein. Although Epstein's model has received a lot of critical acclaim, critics have rejected the idea that anchoring is a metaphysical relation distinct from grounding. I argue that Epstein's grounding-anchoring model is flawed on a more fundamental level than his critics realise. Epstein, who suggests that the grounding-anchoring model is a general model of social kinds, has in fact modelled his account on an erroneous understanding of hybrid kinds. Using the hybrid kind model developed in the previous chapter, I show that the concept of anchoring is ambiguous and that neither Epstein nor his critics have provided an adequate analysis of any of its interpretations.

In the following, I first introduce Epstein's grounding-anchoring model. I then discuss the critics' objections and point out several problems with them. After that, I argue that central examples in the discussion need to be understood as hybrid kinds that beget ambiguity. Once that ambiguity is clarified, it becomes apparent that the grounding-anchoring model cannot be applied to the kinds in question. Turning to Epstein's discussion of non-hybrid kind examples, I show that his model suffers similar problems in this context. I conclude the chapter by providing an error theory of the intuitions motivating the grounding-anchoring model.

2.1 EPSTEIN'S GROUNDING-ANCHORING MODEL OF SOCIAL ONTOLOGY

Epstein's account is motivated by the idea that social ontology can benefit from a closer engagement with existing tools in metaphysics. He suggests that we can advance our understanding of social kinds by distinguishing two metaphysical relations that are at work in the social world: grounding and anchoring (Epstein 2015, 2014). Consider grounding first. The notion of grounding has been developed in recent years in analytical metaphysics and is meant to pick out a specific metaphysical relation (Audi 2012b, Fine 2001, Schaffer 2009, Rosen 2010, Correia & Schnieder 2012).[1] Although there is a fair bit of disagreement about the exact nature of this relation, we can ignore these disputes for most of our discussion and focus on the basic idea instead: the idea that a grounding relation states the "metaphysical" (as opposed to causal) reason why a certain fact obtains. This relation is sometimes described as the "in-virtue-of" relation. A grounded fact, so the basic idea goes, cannot obtain without its grounding fact, unless that grounding fact has been replaced by a different suitable grounding fact. Moreover, some people hold that grounding relations are asymmetric, obtaining between more and less fundamental facts. On this view, the grounds metaphysically explain the grounded facts, but not the other way around. Although Epstein notes that the assumption that some facts about the world are more fundamental than others is highly contested, he explicitly endorses it in the context of his social ontology (Epstein 2015, 70).

A common example of a grounding relation is the relation between biological and physical facts. Biological facts, it is said, are grounded in physical facts, which is to say that biological facts exist *in virtue of* physical facts. According to this idea, a biological fact – such as the fact "hedgehogs have hearts" – cannot exists without its physical grounding fact – such as the arrangements of atoms and molecules that constitute a hedgehog heart – unless the latter are replaced with other grounding facts (imagine a parallel universe where hedgehog have tiny artificial hearts made from plastic and metal).

A grounding relation in Epstein's model obtains between two sets of facts: the *grounds* or grounding facts G, and a specific social fact F. As a toy example, he uses the familiar case of money. Epstein suggests that we could think of the specific social fact "Billy is a dollar bill" as being grounded in the fact "Billy is printed by the Bureau of Engraving and Printing (BEP)". In addition to grounding relations, which apply to singular social fact

[1] Some participants in this debate would deny that grounding picks out a metaphysical relation at all. They suggest that it should be treated as a non-referring expression in the meta-language instead (see Fine 2010).

like "Billy is a dollar bill", Epstein suggests that there are *frame principles*. Frame principles, to put it simple, are generalisations of grounding relations. Instead of saying that fact G grounds fact F ("Billy is printed by the BEP" grounds "Billy is a dollar bill"), they say that facts of type G ground facts of type F (being printed by the BEP grounds being a dollar bill).[2]

In addition to the existing concept of grounding, Epstein proposes a new metaphysical relation of *anchoring*. Epstein introduces the concept of anchoring specifically to further our understanding social ontology, although he suggests in places that it might be applicable to ontological matters more generally. While grounds are the metaphysical reason why a certain social fact or type of social facts obtains, anchors are the metaphysical reason why a certain type of social fact has a specific type of grounding fact (Epstein 2015, 74). Note that an anchoring relation does not hold between the same relata as a grounding relation. A grounding relation holds between a social fact (or a type of social fact) and its grounds. An anchoring relation, by contrast, holds between a frame principle and its *anchors* (or *anchoring facts*). Accordingly, an anchoring relation does not apply to singular social facts but to types of social facts – anchors make it the case that facts of type A ground facts of type B.

Coming back to Searle's analysis of money, Epstein suggests that the grounding-anchoring model applies as follows. As seen in Chapter 1, Searle suggests that the reason bills printed by the BEP are dollars is because we collectively accept that this is the case. According to Epstein, in Searle's example, "x is printed by the BEP" grounds "x is a dollar bill". Furthermore, the fact that this grounding relation obtains is anchored in our collective acceptance of the constitutive rule "bills printed by the BEP are dollar bills". In other words, collective acceptance of the constitutive rule anchors the frame principle "being printed by the BEP grounds being a dollar bill".

To tie the grounding-anchoring model back to our discussion of social kinds, reconsider Epstein's use of frame principles. As argued above, frame principles simply describe grounding relations for types of social facts as opposed to singular social facts. The move from singular facts to types of facts has two upshots. According to Epstein, the main purpose is that it allows us to theorise about counterfactual scenarios – we can say that if the BEP had printed paper bill B (which it did not print in the actual world) B would be a dollar bill. More interestingly for our purposes, it allows us to apply Epstein's model to social kinds. The reason for this is that social kinds can, quite trivially, be

[2] Note that Searle's example refers to *money* rather than *dollar*. The difference, however, is largely immaterial for our purposes.

understood as types of social facts. For instance, the kind *money* is a grouping of all entities x for which "x is money" is true. This implies that there is a straightforward way of talking about the grounds and anchors of a social kind. Social kinds have grounds of the generalising sort expressed in frame principles. For instance, the frame principle ""being printed by the BEP' grounds 'being a dollar"' can be interpreted as saying that "being printed by the BEP" grounds the kind *dollar*. Moreover, since anchors only apply to frame principles (that is, to types of social facts as opposed to singular social facts), it is easy to see how social kinds could have anchors. Anchors are what makes it the case that a social kind has specific membership conditions. On this understanding, to say that collective acceptance of a constitutive rule makes it that case that bills printed by the BEP are dollars is to say that collective acceptance of the constitutive rule anchors the kind *dollar*.

Before we move on to critical replies to Epstein's model, it is important to note the model's intended scope. While it has been widely recognised that Searle's story about collective acceptance applies at best only to a subset of social kinds (see Chapter 1), Epstein proposes his grounding-anchoring model as remarkably more general. His examples include mostly social kinds that would fall under the hybrid model described in the previous chapter such as *law*, *money* and *president*. But Epstein suggests that the model also applies to social kinds like *mob*, which may not fall under the hybrid model because it does not obviously involve a conferred status. In addition, Epstein argues that the model even covers material artefact kinds like *Aldino typeface* and *pocket book* (Epstein 2014).[3] If Epstein succeeded in providing an ontological model that unifies all of these kinds, it would be an impressive achievement.

Unfortunately, there are doubts that his grounding-anchoring model succeeds. There have been a number of critical replies to Epstein's grounding-anchoring model. The most common target of criticism is Epstein's discussion of anchoring. Epstein's description of anchoring largely resorts to metaphors: anchors are facts that "put in place" a frame principle or specific grounding conditions for a social kind (Epstein 2015, 74, 81), or they are the "glue" holding together a social kind (Epstein 2015, 81; 2014). Nevertheless, Epstein insists that grounding and anchoring are different types of metaphysical relations.

[3] Note that material artefact kinds do not fall under my understanding of social kinds. Social kinds, as I understand them, are constituted (partly) by social facts. Members of the kind do not exist in the absence of a social world with certain social relations and institutions in place. Material artefact kinds like *pocket book*, by contrast, depend on social facts in a causal rather than constitutive sense. They were causally produced by humans with a specific purpose in mind, but they do not rely on the existence of society for their own continued existence. If humans went extinct, there would still be pocket books and smartphones, but there would be no hipsters or laws. Material artefact kinds are also not the sort of kinds that typically figure in social scientific explanations, and are therefore of little interest in the context of an investigation that asks if social kinds can be natural kinds.

In particular, he argues against the view that anchors are just among the grounds of social facts, a position that Epstein calls "conjunctivism" (Epstein 2015, Chpt. 9).

Against Epstein's proposal, Jonathan Schaffer and Katherine Hawley defend conjunctivism. They both suggest that anchors are ultimately just grounds for social kinds or facts (Hawley 2017; Schaffer forth.). In other words, Schaffer and Hawley argue that anchors and grounds are the same type of facts and relate to social facts via a grounding relation. Mari Mikkola's takes a slightly different stance (Mikkola 2017). She is willing to accept Epstein's distinction between grounds and anchors to the extent that it allows us to usefully distinguish facts that "work at different ontological levels" (Mikkola 2017, 18). But she rejects the idea that grounding and anchoring are different types of metaphysical relations. According to Mikkola, instead of saying frame principles are *anchored* by a distinct type of anchoring facts, Epstein should say frame principles are *grounded* by these facts on a different ontological level.

To appreciate this subtle distinction, it is important to understand the logical relation between the claims at hand. The claim that anchors are just grounds entails that anchoring is the same as grounding because anchors and grounds are individuated in terms of the metaphysical relation they involve. If all there is are grounds – that is, facts that relate to other, social facts via a relation of grounding – there is no reason to postulate a distinct metaphysical relation of anchoring. According to Mikkola, this is not true vice versa. The idea that anchoring is the same as grounding does not imply that anchors are the same as grounds because the terminology of "grounds" and "anchors" may refer to types of facts that ground social facts in crucially different ways. For the purpose at hand, however, we can largely ignore these differences between Hawley's and Schaffer's criticism on the one hand and Mikkola's on the other. The important upshot is that all three critics take issue with the concept of anchoring, suggesting that it is not a relation sui generis, but is simply grounding in disguise.

In addition to rejecting anchoring as a distinct metaphysical relation, there is another point of overlap in the critics' objections. Both Mikkola and Schaffer object that at least some of the statements Epstein describes as frame principles (that is, statements of generalised grounding relations) do not describe grounding relations at all but are simply identity statements or stipulative definitions. Examples include statements like "committing certain crimes in the context of armed conflict grounds being a war criminal" (Mikkola 2017, 11-13) and "killing someone with deliberate malice aforethought grounds

being a murderer" (Schaffer forth., 14-15). We can ignore these considerations for now, but I will come back to them later in the chapter.

In light of these reactions, it is fair to say that the discussion surrounding Epstein's grounding-anchoring model is far from resolved. It is neither clear whether anchoring is a relation that is genuinely different from grounding, nor whether Epstein's frame principles actually describe an interesting metaphysical relation as opposed to mere identify or definition. Fortunately, the hybrid kind model developed in the last chapter provides a powerful tool for shedding light on these arguments. It allows us to see that both Epstein and his critics mischaracterise the ontology of the social kinds in question. To show this, I first address Epstein's main argument for distinguishing grounding from anchoring. I argue that Epstein and his critics equally fail to recognise that the key examples in Epstein's argument are hybrid kinds which are susceptible to ambiguity. Once this ambiguity is clarified, we can see that Epstein's grounding-anchoring model is inadequate for both hybrid and non-hybrid social kinds.

2.2 THE ARGUMENT FOR DISTINGUISHING ANCHORING FROM GROUNDING

As stated above, a major objection to Epstein's model of social ontology says that anchoring is in fact nothing other than grounding. To evaluate this objection, we first need to consider in more detail Epstein's motivation for distinguishing anchoring from grounding in the first place. For Epstein, the need for a distinct metaphysical relation of anchoring is rooted in the observation that social kinds are "universal tools". Certain facts that are necessary for the existence of a social kind obtain only at the specific time and place that we occupy. Nevertheless, we can reasonably ask whether the kind is instantiated at other times and places. He illustrates this with the example of the kind *war criminal*:

> One is a war criminal if one has committed or conspired to commit any of a long list of crimes in association with armed conflict. We can sensibly ask whether Caligula was a war criminal, or whether Genghis Khan was, having killed over a million inhabitants of a single city. We can also consider a possibility in which some virtuous person instead committed terrible crimes, and sensibly ask whether that person would be a war criminal. It does not matter whether, in that possibility, there is an International Criminal Court. What matters is only whether the person satisfies the conditions we have anchored.
>
> (Epstein 2015, 124)

Let's look at this passage in more detail. Epstein proposes that the grounding facts for *war criminal* are something along the lines of "having committed or conspired to commit certain crimes in association with armed conflict." As a matter of fact, the grounding facts

for *war criminal* may exist or may have existed in many times and places of human history, such as under the empire of Caligula or Genghis Khan. But, Epstein seems to suggests, there is a further crucial ingredient to the kind *war criminal*. In a sense, there cannot be war criminals unless there is an international justice system that decides what it means to be a war criminal – such as the International Criminal Court (ICC).

Curiously, however, it seems that we can reasonably ask whether Caligula or Genghis Khan are war criminals even though the ICC did not exist when they existed. In the passage above, Epstein insists that we can identify members of the kind *war criminal* during the Roman and Mongol empires, long before the ICC was established. This, in Epstein's view, is what justifies the distinction between grounding and anchoring. Grounding is a metaphysical in-virtue-of relation in which both relata need to exist in the same time and place; anchoring is a metaphysical in-virtue-of relation to which this constraint does not apply.

To clarify this point, Epstein argues that "when we assess [social facts] across other times and possibilities, we do not deny that they obtain merely because the anchoring facts do not obtain at those times and possibilities." (Epstein 2015, 124). In other words, he suggests that the grounding relation described in the frame principle ("committing or conspiring to commit certain crimes in the context of armed conflict grounds being a war criminal") obtains even in contexts where the anchors for the frame principle do not exist. Instead, when asking whether Genghis Khan was a war criminal, we can simply evaluate the situation in the light of the anchors that obtain at the time and place from where we ask the question. According to Epstein, we can only make sense of this observation if we assume that anchoring and grounding are different types of relations.

Epstein's argument for this claim is a bit elliptic, and he does not discuss the anchors for *war criminal* in any detail other than suggesting they involve the ICC. For our purposes, it helps to be a little more precise. We can say that *war criminal* is anchored by certain facts about the ICC, such as that the ICC exists and that it adopted specific legal instruments concerning war crimes, such as the Rome Statute. Accordingly, Epstein's argument for distinguishing grounding from anchoring seems to go something like this:

(1) Per definition of grounding, a social kind can only be instantiated when its grounds obtain.
(2) The anchors for *war criminal* (the fact that the ICC exists, the fact that the ICC adopted the Rome Statute, etc.) did not obtain when Genghis Khan existed.
(3) Genghis Khan was a war criminal, that is, he instantiated the kind *war criminal*.

(4) Ergo: There must be a distinct relation (called "anchoring") which holds between a social kind like *war criminal* and its anchors.

Epstein's argument has attracted a number of critical replies. Interestingly, the objections are quite diverse and some of them are straightforwardly inconsistent with each other. Consider Hawley's argument first. Hawley challenges premise (3). She denies that it makes sense to say Genghis Khan was a war criminal, at least in some contexts:

> […] there are contexts in which it is not sensible to ask whether [Genghis Khan] was a war criminal, unless we mean to ask whether he violated whatever laws may have governed warfare at his time. Such imposition of modern categories onto historical figures is regarded by many (though not all) historians as anachronistic, and explanatorily fruitless. Thus, socially oriented historians of science, and sociologists of contemporary science, stress the importance of using "actors' categories", i.e. concepts available to the scientists [sic., actors] who are the subjects of study.
>
> (Hawley 2017, 10)

In other words, Hawley objects that it is wrong to say that anchors generally behave differently to grounds. She points out that there are social scientific inquiries in which anchors play much the same role as grounds. Just as it would be wrong to classify Genghis Khans as a war criminal had he not committed or conspired to commit any crimes, social scientists might argue that we cannot say that Genghis Khan was a war criminal unless an equivalent of the ICC existed at the time when he committed his atrocities.

Schaffer makes a very similar point (Schaffer forthc.). He argues that it is obvious that anchors behave like grounds when we consider the example of *money*. To see this, Schaffer asks us to imagine a scenario where the United States adopted Spanish dollars instead of U.S. dollars as their currency. The laws that entitle the BEP to print currency have not been passed, but there exists some alternative legislation regarding Spanish dollars. The BEP is still in the business of printing paper notes, but the notes are parking tickets rather than money. In this scenario, Sally holds a Spanish dollar in her right hand and a parking ticket in her left hand. We need to decide in which hand Sally is holding her money. According to Schaffer, Epstein's grounding-anchoring model would commit him to saying the money is in Sally's left hand. The reason for that, he suggests, is that Epstein's model states that we need to answer the question based on which anchors obtain in the context from where we ask the question. In the actual world, there exists U.S. legislation which anchors the frame principle that being printed by the BEP grounds being money. Therefore, Schaffer concludes, Epstein would be committed to saying that the piece of paper in Sally's left hand is money. That, Schaffer objects, is implausible – the sensible

thing to say would be that Sally is holding money in her right hand. According to Schaffer, whether the piece of paper in Sally's right hand is money depends on what anchors obtain in the hypothetical scenario, for instance whether there exists legislation that makes Spanish dollars money. Just as with grounds, what matters are the anchors for money that obtain in the hypothetical scenario. As a result, Schaffer concludes that Epstein's anchors function in just the same way as grounds.[4]

Interestingly, while Hawley and Schaffer make a similar general point – that there is reason to think that anchors are just grounds – they disagree quite substantially on the nature of a specific case study. Hawley, as argued above, accepts that the statement "committing or conspiring to commit certain crimes in the context of armed conflict grounds being a war criminal" expresses a genuine grounding relation but denies that we can unambiguously classify Genghis Khan as a war criminal. Schaffer takes no issue with classifying Genghis Khan as a war criminal, but denies that "committing or conspiring to commit certain crimes within the context of armed conflict" are the grounds of the kind *war criminal*. According to Schaffer, the reason we can call Genghis Khan a war criminal is because "war criminal" is stipulatively defined as someone who commits or conspired to commits certain crimes in the context of armed conflict. In other words, "having committed or conspired to commit certain crimes in the context of armed conflict" is simply an explication of how we have stipulatively defined the kind that "war criminal".

This has crucial implications on how we understand the role of the ICC in relation to *war criminal*. According to Schaffer, facts about the ICC are neither grounds nor anchors for *war criminal*. Instead, they are simply historic facts about how (when, where, by whom, etc.) the term "war criminal" has been stipulatively defined. Once stipulated, definitions apply universally. We can correctly use a term to describe an individual even when facts about how the term has been defined no longer obtain.

Incidentally, Schaffer is not alone with the idea that some of Epstein's examples of grounding relations are simply definitions. With regard to Epstein's example of murderer, Mikkola argues that it is doubtful whether being a *first-degree murderer* is really grounded by killing someone with deliberately premeditated malice aforethought. According to Mikkola, just like a square is an equilateral right quadrilateral, "being a murderer just is to kill with premeditation" (Mikkola 2017, 11-12). However, unlike Schaffer, Mikkola argues that the fact that these statements express definitions does not automatically imply that

[4] Epstein would probably reject this argument because he wants to treat counterfactual scenarios like the money example differently from countertemporal scenario like the one involving Genghis Khan. But, as noted by Hawley (2017, 9-10), it is not clear on what basis he could make such a distinction.

they cannot be grounding statements. That, Mikkola argues further, depends on whether they are identity statements. Whether or not the statements in questions are identity statements in turn depends on how we individuate facts. For Epstein, who thinks that we should individuate facts by the propositions that they express, definitions can involve grounding relations. On a propositional conception of facts, statements like "x is a square" and "x is an equilateral right quadrilateral" describe two different facts. Hence, statements like "squares are equilateral rights quadrilaterals" are not identity statements and can express grounding relations.

However, Mikkola points out that this would not be the case on a "worldly" conception of facts, where facts are "individuated by their constituents and the manner in which those constituents are combined" (Audi 2012a, 103; cited in Mikkola 2017). On a worldly conception of facts, the statements in question are indeed identity statements that cannot express grounding relations. According to Mikkola, Epstein fails to provide an argument as to why a propositional conception of facts is more suitable to the purpose of social ontology than the worldly conception. In other words, unlike Schaffer, Mikkola suggests that definitions can express grounding relations in Epstein's account. But she criticises that the conception of facts which Epstein uses to make this possible lacks independent justification.

The replies to Epstein's argument are puzzling in several ways. For one thing, some of them – such as Hawley's and Schaffer's interpretations of the example of war criminal – are obviously in tension with each other. For another, there are problems with each of the arguments taken by themselves. Hawley seems to suggest that there might be some contexts where it makes sense to say "Genghis Khan was a war criminal" and other contexts where it does not makes sense, but she does not tell us what distinguishes the former from the later. Schaffer claims that some of Epstein's examples are cases of grounding whereas others are stipulative definitions, but he does not clarify how exactly we can tell the former from the latter. The same is true of Mikkola's suggestion that, on a "worldly" conception of facts, some of Epstein's examples are identity statements.

Overall, then, the critics' responses to Epstein's distinction between anchoring and grounding pose more questions than they answer. The reason for that, I suggest, is that both Epstein and his critics fail to fully recognise the ontological structure of the examples they are discussing. Kinds like *money* or *war criminals* are hybrid kinds. Since hybrid kinds consist of a base kind and a status kind, identifying the grounds and anchors of a hybrid kind requires substantial disambiguation. In the following, I disambiguate Epstein's claims

about grounding and anchoring and explore what implications this has for his model of social ontology.

2.3 GROUNDING AND ANCHORING FOR HYBRID KINDS

2.3.1 Was Genghis Khan a war criminal?

Let's begin by reconsidering the central example of *war criminal*. Schaffer suggests that a war criminal is someone who commits specific crimes in the context of armed conflict, as defined by the ICC. Furthermore, he argues that this answer holds for the Mongol Empire just as much as for today. Hawley, by contrast, argues that it is sensible to say that a war criminal is someone who violates whatever laws govern warfare in their specific time and place.

Who is right? The hybrid kind model developed in Chapter 1 provides a useful tool for resolving this discussion. The model suggests that some social kinds are best understood as hybrid kinds, consisting of a base kind and a status kind connected by conventional linkage. *War criminal* can be understood as a hybrid kind. The base kind of *war criminal* is the kind *people who commit or conspire to commit certain acts in the context of armed conflict*. It is this kind that we try to track when applying the classification of *war criminal*. But the story does not end here. Once we have identified a person who we believe has committed or conspired to commit the relevant crimes, we apply the classification of *war criminal* to them. By doing so, we confer onto the individual a specific social status that is constituted by a number of legal and social properties (such as being hunted by the police, being condemned by the public, etc.).

This characterisation, of course, is a highly simplified version of how the process of status conferral actually works in this case. In reality, the process encompasses a complex nested structure of status conferrals involving the ICC and numerous other institutions. Nevertheless, the example allows us to realise that the relationship between base kind and status kind is one of conventional linkage by recognising how easily the two can come apart. The ICC might not be perfect – it may fail to commit the status onto individuals who have committed the relevant acts, and it might erroneously confer the status onto individuals who have not committed or conspired to commit the relevant acts. In such cases, an individual might have committed or conspired to commit certain crimes in the context of armed conflict but not occupy the legal and social position of a war criminal, or vice versa.

Understanding *war criminal* as a hybrid kind allows us to see that the discussion between Hawley and Schaffer has fallen victim to a common confusion with hybrid kind terminology. Since a hybrid kind consists of a base kind and a status kind which are connected in a peculiar way, the terms used to refer to these kinds are highly susceptible to ambiguity. Against this background, we can see that Hawley's and Schaffer's different linguistic intuitions are both warranted. Recall that, for Schaffer, it makes perfect sense to say Genghis Khan was a war criminal, because a war criminal simply *is* someone who, as defined by the ICC, committed or conspired to commit specific crimes in the context of armed conflict. Hawley, by contrast, has reservations about applying the term "war criminal" to Genghis Khan. Echoing the concerns of certain historians and sociologists, she suggests that calling Genghis Khan a war criminal makes sense only when we mean to say that Genghis Khan violated laws that governed warfare at his time.

We can console these two seemingly conflicting views once we recognise that "war criminal" refers to a hybrid kind and therefore is ambiguous. As argued in Chapter 1, the hybrid kind term "money" can refer either to the status kind (being a means of exchange, etc.), to the base kind (specific pieces of paper), or to the hybrid kind as a whole. In the same way, "war criminal" can refer either to a base kind (people who commit or conspire to commit certain crimes in the context of armed conflict), to the status kind (people who occupy the legal and social position of someone classified as a war criminal) or the entire hybrid kind.

On this view, we can see that Schaffer's and Hawley's claims are both correct. Schaffer, who suggests that the meaning of the term "war criminal" is someone who commits or conspires to commit certain crimes in the context of armed conflict, interprets "war criminal" in terms of its base kind. An interpretation of "war criminal" in terms of the base kind supports Schaffer's claim that the term applies to Genghis Khan. Anyone who committed or conspired to commit certain crimes in the context of armed conflict is a war criminal on this interpretation.

Hawley's use of the term "war criminal", by contrast, is best understood as referring to the status kind – the specific legal and social position that is conferred onto someone who is believed to have committed or conspired to commit certain crimes in the context of armed conflict. This interpretation supports Hawley's suggestion that Genghis Khan was not a war criminal. When we use the term "war criminal" to refer to a status kind, the status kind we have in mind is a product of modern legislation and social practice. Whatever laws and public understanding were associated with warfare at the time of the

Mongolian empire, they were probably quite different from today's. Therefore, even if a classification akin to *war criminal* existed at the time, it probably carried a different legal and social meaning. In other words, although the acts committed by Genghis Khan might have been regarded as cruel and atrocious in his time, it is very unlikely that they had the same legal and social status that they would have today. Against this background, Hawley's suggestion that calling Genghis Khan a "war criminal" might be anachronist and explanatorily fruitless makes sense. Since Genghis Khan did not occupy the relevant legal and social position that characterises the status kind associated with *war criminal*, it would be wrong to apply the term to him in that interpretation.

2.3.2 Implications for the grounding-anchoring model

In the previous section, I argued that understanding *war criminal* as a hybrid kind allows us to realise that Hawley and Schaffer both provide correct yet incomplete analyses of the phenomenon in question. It is now time to consider how these findings square with Epstein's grounding-anchoring model. As suggested above, applying Epstein's grounding-anchoring model to the kind *war criminal* gives the following analysis: The frame principle for war criminal is "committing or conspiring to commit certain acts in the context of armed conflict grounds being a war criminal". This frame principle is anchored, among other things, by the ICC's adoption of statutes which say that someone who commits or conspires to commit certain act in the context of armed conflict is a war criminal, and ought to be legally punished in a certain way.

The problem with this analysis is that it obviously ignores the hybrid nature of *war criminal* described above. Since *war criminal* is a hybrid of two kinds, we would expect at least two sets of frame principle and anchoring relation, one for the base kind and one for the status kind. Let's see how this might work out. To avoid confusion, I will use "war criminal$_B$" to refer to the base kind and "war criminal$_S$" to refer to the status kind.

Consider the base kind first. The corresponding frame principle would state "committing or conspiring to commit certain acts in the context of armed conflict grounds being a member of *war criminal*$_B$". This is odd. As argued above, the base kind for war criminal just is the grouping of people who commit or conspire to commit certain crimes in the context of armed conflict. In other words, when we talk about *war criminal*$_B$, Schaffer's and Mikkola's objections hit a nerve. On this interpretation, the frame principle simply describes how "war criminal" has been stipulatively defined. As Mikkola has argued, whether such definitions can express grounding relations depends on which conception of facts one adopts in the context of social ontology. However, we can take

her argument somewhat further. While there is some discussion as to whether or not real definitions describe grounding relations (e.g. Rosen 2015), I am not aware of any such arguments with respect to stipulative definition. In any case, it is hard to see how stipulative definitions could be non-symmetrical, or how they could involve the lower-level and higher-level facts that Epstein suggests are required for grounding. As a result, even if the frame principle for *war criminal*$_B$ expresses a genuine grounding relation, it is at best a quite untypical examples of the concept of grounding that Epstein has in mind.

Likewise with the anchors for *war criminal*$_B$. If we accept that the frame principle above expresses a genuine grounding relation, the corresponding example of anchoring looks very peculiar. To say that the ICC's decision to call people who commit or conspire to commit certain crimes in the context of armed conflict "war criminal" anchors *war criminal*$_B$ is akin to saying that my decision to define "splott" as the set of items in the bottom of my freezer anchors *splott*. The question why we use certain terms to refer to specific concepts or kinds might of interest to linguists. But it is hard to see how it could involve a distinct metaphysical relation that has previously gone unnoticed and warrants the introduction of a novel ontological term.

In other words, if we interpret "war criminal" in terms of the base kind, Schaffer's and Mikkola's objections are correct. The frame principle states a definition, and the alleged anchors are facts about how the definition has been put into place. This shows that the grounding-anchoring model is inadequate for the purpose at hand. It does not provide any novel or plausible insights about the ontology of *war criminal*$_B$. At best, it expresses in a more convoluted way relations which we already understand pretty well.

Reason enough to move on to the next candidate. What would a grounding-anchoring analysis of the status kind *war criminal*$_S$ look like? In this case, the frame principle would be "committing or conspiring to commit certain acts in the context of armed conflict grounds being a member of *war criminals*$_S$". There is some reason to think Epstein intends for his model to apply to status kinds like *war criminal*$_S$. To see this, consider his discussion of the example of *first-degree murder*. Epstein's discussion of *first-degree murder* is based on a paragraph from the Massachusetts General Law (MGL), which says:

> Murder committed with deliberately premeditated malice aforethought, or with extreme atrocity or cruelty, or in the commission or attempted commission of a crime punishable with death or imprisonment for life, is murder in the first degree.
>
> (MGL c. 265 §1; cited in Epstein 2015, 89)

Anticipating Mikkola's concerns, Epstein acknowledges that this paragraph, which states that murder committed with deliberately premediated malice aforethought *is* first-degree murder, might look like an identity statement. But he insists that this first impression is mistaken, and that the paragraph is better understood as expressing a grounding relation. To support this idea, Epstein refers to Searle's discussion of social kinds, arguing that "Searle rightly stresses that being a first-degree murderer is a *status*. Statuses are not identical to their conditions. Being a dollar is not the same thing as being printed at a certain bureau." (Epstein 2015, 92, my emphasis). This passage strongly suggests that Epstein's grounding-anchoring model should be understood as a model of status kinds. Interpreted in this way, Epstein is no longer susceptible to the criticism that his frame principles are just identity statements or stipulative definitions – committing or conspiring to commit certain crimes in the context of armed conflict is clearly not the same as occupying a specific social and legal status.

The resort to status kinds, however, comes at a high cost because now a different problem arises. While the frame principle, interpreted as a claim about a status kind, does not state a stipulative definition of *war criminal*$_S$, it does not describe a grounding relation either. As argued above, the relation that connects "committing or conspiring to commit certain crimes within the context of armed conflict" with the status kind of war criminal is the relation of status conferral. The status of *war criminal* is conferred onto individuals which are believed to have committed or conspired to commit the relevant crimes.

Status conferral is evidently different from grounding. In the previous chapter, I suggested that Asta's choice to call base properties "grounding properties" is a misnomer. Now that we have a better grasp on the notion of grounding, it is easy to see why. Grounds are the metaphysical reason (or part of the metaphysical reason) why a certain fact obtains. In the example at hand, this means the grounds of *war criminal*$_S$ are the metaphysical reason (or part of the metaphysical reason) why someone is a member of the status kind *war criminal*$_S$. But the fact that someone has committed or conspired to commit certain crimes in the context of armed conflict is clearly not the metaphysical reason (or part thereof) why that person is a member of *war criminal*$_S$. Being a member of *war criminal*$_S$ means occupying a specific legal and social position. To occupy that position, it is quite immaterial whether or not one actually has committed or conspired to commit certain acts in the context of armed conflict. What matters is that people (and, in a derivative sense, institutions like the ICC) believe one has committed or conspired to

commit these acts, and that they change the way they think about and interact with the individual accordingly.

If anything, then, the role of the alleged grounds (committing or conspiring to commit certain crimes in the context of armed conflict) is more akin to a causal one. Quite plausibly, the fact that someone has committed or conspired to commit certain crimes in the context of armed conflict tend to play a causal role in leading the ICC to believe that the individual has committed or conspired to commit certain acts in the context of armed conflict. This belief, in turn, may cause a range of people to interact with and think about the individual in a way that makes him or her occupy the legal and social position of *war criminal*$_S$.[5] Since the relation in question is not one of grounding, Epstein's concept of anchoring – the metaphysical reason why a social kind has specific grounds – cannot be applied to *war criminal*$_S$ either.

In other words, Epstein's grounding-anchoring model cannot be usefully applied to the hybrid kind *war criminal*. The suggested frame principle either describes a stipulative definition (in the case of *war criminal*$_B$), which is at best a questionable candidate for grounding. Or, when applied to *war criminal*$_S$, it describes a relation of status conferral, which is clearly different from grounding. This has an impact on Epstein's idea of anchoring. Since the frame principle for *war criminal*$_S$ clearly does not express a grounding relation, the concept of anchoring cannot be applied. In the case of *war criminal*$_B$, where the frame principle might be interpreted as stating a (untypical) grounding relation, the alleged anchors are simply facts about how a definition has been put into place.

Does this mean that Epstein's model is unsuitable for hybrid kinds in general? Not necessarily. There is still the off-chance that cases like *war criminal* (and maybe *first-degree murder*) are just idiosyncratic examples. To address this concern, it is worth briefly considering another example, the classic case of *dollar*. The frame principle for *dollar* states "being a bill printed by the BEP grounds being a dollar". According to Schaffer, unlike with the problematic cases discussed above, the frame principle for *dollar* involves a genuine grounding relation.

To see whether the grounding-anchoring model can be successfully applied to the hybrid kind *dollar*, we again need to start by disambiguating the terminology. As argued before, the base kind of *dollar* consists of bills printed by the BEP, with the associated status kind being a means of exchange. Once the term "dollar" is disambiguated, it is easy

[5] In other words, status conferral can be considered a causal relation if one believes in mental causation. Asta, by contrast, seems to suggest that it is a relation sui generis (e.g. Asta 2008). My thesis does not take a stand on this issue.

to see how Epstein's model runs into the same problems. On the base kind interpretation of "dollar", the frame principle again simply describes what has been defined by stipulation – that the base kind for *dollar* are bills printed by the BEP.[6] If we interpret "dollar" as a status kind instead, the frame principle effectively says "being a bill printed by the BEP grounds being a means of exchange". Again, this statement does not describe a grounding relation, but a relation of status conferral. As argued in Chapter 1, bills printed by the BEP function as a means of exchange because we confer that status onto them.

In other words, contrary to Schaffer's suggestion, the frame principle for *dollar* does not involve a grounding relation that is genuinely different from definition. Instead, the example runs into exactly the same problems as those we encountered with *war criminal*. This supports the idea that the grounding-anchoring model cannot be usefully applied to hybrid kinds in general. Furthermore, it allows us to see that both Epstein and his critics are wrong about the notion of anchoring. Anchoring is not a metaphysical in-virtue-of relation that is distinct from grounding, but anchoring is not grounding either.

2.4 ANCHORING IN NON-HYBRID KINDS

2.4.1 Anchoring as "gluing"

While we have established that Epstein's model is not suited for hybrid kinds, including many of his central case studies, it is too early to reject the model altogether. After all, Epstein is not the first to trip up on the phenomenon of hybrid kinds, and there a many other, non-hybrid kinds for which the grounding-anchoring model might be apt. In order to explore this option, we need to move away from the problematic hybrid kind examples and consider what Epstein has to say about grounding and anchoring more generally.

A fruitful place to begin is by reconsidering the intended scope of Epstein's model. Up until now, we have assumed that the grounding-anchoring model is limited to social kinds. In a remarkable passage, however, Epstein suggests that the scope of the model in principle extends to natural kinds as well. Referring to anchoring in terms of the metaphor of "gluing", Epstein suggests that

> It is a general feature of kinds—not just social kinds like dollars and play tea parties—that something needs to *glue* them together. Even a natural kind like gold may need a bit of "glue," to set it up as a natural kind. Some philosophers hold, for instance, that laws of nature play some role in acting as this glue. The idea is that all it takes for an object to be a sample of

[6] This is not to say that all base kinds of hybrid kinds are established by stipulation. Some are established by social practices following implicit rules, as might be the case for gender and race in some contexts. In these cases, an explicit definition of the base kind would have more of the explanatory asymmetry associated with real definitions, and hence would be a more plausible candidate for grounding. For reasons I explain in the next section, this does not mean that the grounding-anchoring model provides and adequate analysis of these kinds.

> gold is to be composed of atoms with a particular atomic number. However, what unifies a chemical kind (like gold) into a natural kind is that the laws of nature make the chemical behave in certain regular ways. Without laws gluing the chemical kind together, it would not be a natural kind at all.
>
> (Epstein 2015, 81, original emphasis)

This passage is helpful in several ways. Firstly, it makes clear that Epstein believes anchoring holds for both natural and social kinds. Secondly, the example of gold gives us a more detailed understanding of what anchors might be.

Let's look at this in more detail. As Epstein (2015, 67-68) acknowledges, a common way of characterising kinds (both natural and social) is by contrasting them with ad hoc sets of things. Members of kinds (such as *tiger* or *communist*) are said to have a great number of properties in common, whereas members of ad hoc sets of things (such as *white things* or *things I can see from my desk*) do not. This characteristic is not only used to distinguishing kinds from ad hoc sets of things. It is also assumed to explain the central role that kinds play in science.[7] The fact that members of a kind have a great many properties in common, so the idea, facilitates induction. It allows us to infer from the fact that an entity has some property (such as having atomic number 79), to the fact that is has a great number of other properties (such as being yellow, malleable, and ductile).

Coming back to Epstein's passage above, this understanding of kinds suggests that Epstein's metaphor or "gluing" hints at whatever makes it the case that the properties associated with a kind are reliably coinstantiated. To avoid misunderstanding, the underlying idea here is not that the properties of individual objects are at the risk of running off into all directions unless they are firmly held together. Instead, it is the assumption that there must be arrangements in the world which make it the case that objects which exemplify some properties associated with a specific kind will also exemplify other properties associated with the kind.

Epstein points out that, for natural kinds, it is typically assumed that this function is carried out by laws of nature. In a footnote, he qualifies this claim by acknowledging that there are other candidates for bringing about this stable association of properties (Epstein 2015, 81n5). Although Epstein does not mention it in this context, another such prominent "glue" candidate for natural kinds are homeostatic mechanisms (Boyd 1991, 1999; Millikan 1999). According to Boyd's homeostatic property cluster (HPC) account,

[7] I will later argue that this conception of (natural) kinds is overly simplistic (see Chapter 4). However, we can ignore these problems for now as they are not relevant to the discussion at hand.

natural kinds are constituted by clusters of properties which are reliable coninstantited in virtue of specific causal factors or mechanisms.

Although the passage above only mentions natural laws as potential anchors for natural kinds, in a different paper Epstein discusses homeostatic mechanism as potential anchors for social kinds (Epstein 2014). The aim of this paper, according to Epstein, is to show that social kinds can have diverse anchors (or "glues") holding them together. To illustrate this, Epstein discusses the (allegedly) social kinds *italic, Aldino typeface,* and *pocketbook*. According to Epstein, all three have slightly different types of grounds. While grounds for *italic* are purely qualitative (to be italic, a font merely needs to have the qualitative feature of being written in a slanted style), the grounds for *Aldino typeface* are qualitative and "historical" and the grounds for pocketbook are qualitative and "functional". In order to be *Aldino typeface*, a font not only needs to have certain qualitative features, it also needs to be a copy of Griffo's original punches of Aldino typeface. And in order to be a *pocketbook*, an object not only needs to have the qualitative features of a book, it also needs to have the function of being easily carried around in a pocket.

For the purpose at hand, we can put aside the question whether Epstein has accurately identified the grounds for these kinds. What is important is his discussion of anchoring that follows. According to Epstein, the reason *Aldino typeface, pocketbook* and *italic* have such different types of grounds lies with the fact that they have different types of anchors. Epstein's reasoning here is somewhat fragmentary, but he seems to suggest that *Aldino typeface* is anchored as a historical kind in the sense developed by Ruth Millikan (Millikan 1984, 1999). *Aldino typeface's* "dominant glue", Epstein suggests, is "the functional explanation for the proliferation of that particular family" (Epstein 2014, 14). In other words, what anchors *Aldino typeface* is a mechanism of (evolutionary) function – the reason that the properties of *Aldino typeface* are reliable associated is because the properties perform a specific function that allows them to be propagated through copying. Since, on this view, the copying of instances with specific qualitative properties is tightly linked with family lineages, Epstein concludes that the grounds for *Aldino typeface* involve family membership.

The anchors for *pocketbook*, Epstein suggests, are slightly different. As with *Aldino typeface*, Epstein suggests that the reason that entities with the qualitative characteristics of pocketbooks exist is because they perform a certain function. However, unlike with *Aldino typface*, having the relevant *pocketbook* properties is not intricately linked to a specific lineage. Instead, *pocketbooks* can belong to a number of different copying lineages. As a

result, the grounds for *pocketbook* involve performing a specific function, but not membership in a specific lineage.

In both cases, Epstein's account of anchoring invokes the idea that specific property clusters are reliably coinstantiated because the properties fulfil specific functions. The parallels with HPC accounts of biological species are obvious and intentional. In a footnote, Epstein acknowledges that his image of "gluing together" a kind is taken from an exchange between Ruth Millikan and Richard Boyd on precisely this topic (Boyd 1999, Millikan 1999). More than that, he suggests that the theories discussed in this exchange "are, in part, theories of anchoring schemas" (Epstein 2014, 11n16).

We need to be careful when interpreting this claim. At first glance, it may seem that Epstein is suggesting that "gluing" is the same as anchoring. This would directly contradicts Epstein's core assumption that anchoring is non-causal relation. As explained above, the HPC account is an account of causal mechanisms. Homeostatic mechanisms "glue together" a kind in the sense that they cause properties to be reliably coinstantiated. In other words, "gluing" holds together the properties that constitute a kind. This is not the same as anchoring. Anchoring, according to Epstein, is a relation that holds together a social kind (or type of social fact) and its grounds. What Epsteins seems to be saying, then, is that "glues" are anchors. Social kinds are anchored by the existence of specific "glues", that is, homeostatic mechanisms that cause properties to be reliably coinstantiated. The existence of these mechanisms is what makes it the case that a social kind k has certain grounds G.

2.4.2 Problems with anchoring as "gluing"

Epstein's discussion of anchoring is odd for several reason. Firstly, there is the fact that all three examples are artefact kinds. As mentioned at the outset, artefact kinds are somewhat peculiar examples of social kinds in that they (i) depend on human social life primarily in a causal rather than a constitutive sense, hence do not share what some regard the core ontological feature of social kinds; and that they (ii) are not generally the sort of kinds that play a central role in social scientific inquiry, hence are not the most obvious analogues of natural kinds. I suggest that we put this issue aside for now. If Epstein's account proves to be promising with respect to artefact kinds, we can still worry about whether it can be generalised to more paradigmatic social kinds. Secondly, there is the question to what extent evolutionary functions can be attributed to cultural artefacts. There is a lively debate on this topic that I will not attempt to resolve here (Lewontin & Faccia 1999; Kronfelder 2010, Lewens 2015, Mesoudi et al. 2006). Instead, I will give

Epstein the benefit of the doubt and assume that the models can be applied to the examples he uses.

Still, there remain further concerns which we cannot put aside so easily. These concerns are familiar from the discussion of hybrid kind examples above, and symptomatic of a more general problem with Epstein's account. In the section on hybrid kinds, we saw that interpreting frame principles like "being a bill printed by the BEP grounds being a dollar" as referring to the base kinds of *dollar* implies that the frame principles state stipulative definitions. Now, we can see that this phenomenon is not an anomaly that occurs when we apply the grounding-anchoring model to hybrid kinds, but affects non-hybrid kinds as well. Epstein's presentation of non-hybrid examples suggests that the frame principles have the form "Having property cluster G grounds being a member of kind k". Depending on the example, property cluster G is either family membership plus qualitative properties, function plus qualitative properties, or merely qualitative properties. But by saying this, Epstein effectively suggests that the grounds for kind k are simply k's instantiation conditions (see Epstein 2014, 3n4). As a result, applying the grounding-anchoring model runs into similar problems to the one's discussed in the context of base kinds. The resulting frame principles describe definitions, and whether or not these involve genuine grounding relations depends on one's conception of facts.

But the problems go further still. Even if we admit that frame principles of the form "having property cluster G grounds being a member of k" describe a genuine grounding relation, it would not be the type of grounding relation that could possibly be anchored by homeostatic mechanisms. Anchors, as is clear from Epstein's characterisation, are supposed to be facts which make it the case that a specific type of fact is grounded in a specific other type of fact. With respect to kinds, as argued above, this is translated into anchors being the facts that make it the case that a kind k has specific instantiation conditions G (rather than G*). In other words, anchors are what makes it the case that having property cluster G (rather than G*) grounds being a member of kind k. But the existence of a homeostatic mechanism simply does not make it the case that having property cluster G, rather than G*, grounds being a member of kind k. If anything, the existence of a homeostatic mechanism makes it the case that having property cluster G grounds being a member of a *kind* as opposed to an ad hoc collection of properties. If the homeostatic mechanisms were to disappear, the grounds for k would not suddenly change from G to G*. Instead, k would still be characterised (or grounded) by the very same properties, but it would cease to be a kind. The reason for that is that there is nothing in

the world that guarantees that properties G will be reliably coinstantiated. As a result, k loses its ability to facilitate inductive inferences. It enters the humble ranks of ad hoc collection of properties instead. In other words, Epstein's reference to homeostatic mechanisms fails to elucidate his notion of anchoring. We are still left with an abstract, obscure concept that does not advance our understanding of social ontology.

2.5 ANCHOR AWEIGH – AN ERROR THEORY OF THE GROUNDING-ANCHORING MODEL

The discussion above showed that the problems with Epstein's grounding-anchoring model are not limited to social kinds, but affect all kinds – natural and social, hybrid and non-hybrid – alike. How could Epstein's account go so wrong? The purpose of this question is not to deride Epstein's attempt at providing unified social ontology. As the discussion of the critical replies to Epstein's model suggest, he is in excellent company. The fundamental flaws that underlie his grounding-anchoring model were either unnoticed or implicitly or explicitly shared by his critics. The reason for this, I believe, is that the problems with Epstein's model reflect a widespread misunderstanding of social kinds which leads philosophers to incorrectly conceptualise hybrid social kinds and to fail to distinguish hybrid from non-hybrid kinds. In the remainder of the paper, I will sketch an "error theory" for the grounding-anchoring model which is based on this observation.

To see just how easily the mistakes are made that trouble both Epstein and his critics, imagine starting to explore the topic of social ontology with an example of a non-hybrid social kind, such as *inflation*. Inflation is the process by which the general prices for goods and services are rising, and, conversely, by which the purchasing power of a currency is falling. In other words, the instantiation conditions for *inflation* are "being a process by which the general prices for goods and services are rising, and, conversely, by which the purchasing power of a currency is falling". Now, consider the question of anchors, i.e. ask yourself "what could possibly make it the case that these are the instantiation conditions for *inflation*?" If your linguistic intuitions are the same as mine, this question looks straightforwardly odd. Giving it some further thought, you might feel inclined to say the question is pointless because "inflation" is simply defined as the process by which the general prices for goods and services are rising, and, conversely, by which the purchasing power of a currency is falling

However, when you begin your exploration with the example of *money* (or *dollar*), which for historic reasons has become the paradigmatic example of a social kind, the situation

looks very different, because it also happens to be a paradigmatic hybrid kind. The associated statement that "bills printed by the BEP are money" is not obviously a definition or an identity statement. After all, in other societies, it is not bills printed by the BEP that are money, but rather bills printed by the Central Bank, or shells, or cigarettes. As a result, it suddenly seems to make perfect sense to ask "what makes it the case that these are the instantiation conditions for *money*?" In fact, Searle himself suggested that there must be a non-causal, metaphysical relation connecting these different types of facts (Searle 1995, 42). We know now that this impression results from the hybrid structure of the kind *money*, meaning that the term "money" is ambiguous and may refer both to the base kind (bills printed by the BEP) and to the status kind of money (a means of exchange). The problem underlying Epstein's grounding-anchoring model, as well as his critics' responses, is that they are informed by linguistic intuitions about statements whose ambiguity has hitherto gone unnoticed in social ontology.

2.6 CONCLUSION

In this chapter, I argued that Epstein's grounding-anchoring model of social ontology is conceptually flawed in a way that extant criticism of his model has failed to pin down. The reason for this is that paradigm examples of Epstein's account are hybrid kinds which beget terminological ambiguity. Disambiguation of the relevant kind terms suggested that Epstein's examples of grounding and anchoring are not what he makes them out to be. Alleged grounding relations turned out to be either definitions or relations of status conferral. Alleged anchoring relations are neither grounding relations nor non-causal metaphysical relations distinct from grounding. Instead, anchoring can either be understood as a relation of how a definition has been put in place, or as a causal relation whereby anchors make it the case that a certain status is conferred onto objects of a base kind. As a result, I argued, the grounding-anchoring model is not applicable to the hybrid kind cases that are used in the discussion.

I then showed that the problems with the grounding-anchoring model are not limited to hybrid kinds. The frame principle for non-hybrid kinds, too, describe definitions rather than typical examples of grounding relations. Furthermore, the analysis of anchors as homeostatic mechanism that Epstein offered in the context of non-hybrid kinds is implausible. While homeostatic mechanisms may stabilise property clusters in a way that makes them kinds rather than ad hoc collections of things, anchors are not responsible for a kind having specific instantiation conditions. In the penultimate section, I suggested

that the reason Epstein's model faces so many difficulties is because the central idea of anchoring is incoherent. Unless it plays on the ambiguity of a hybrid kind term, the question why a kind k has instantiation conditions G is not an interesting or meaningful question in the first place. Having clarified the ontological structure of hybrid kinds, as well as the confusion it has caused in central debates on the nature of social ontology, it is time to turn to a different set of questions. In the next chapter, I consider what the hybrid model can tell us about the role of moral and political values in social ontology.

3

VALUES IN SOCIAL ONTOLOGY

The discussion on the role of moral or political values in social ontology is a polarised one. Some people insist social ontology has no room for these values. If the social sciences want to make sense of the social world, the kind concepts that they use should simply reflect the way the world *is*. Moral and political values ought to be kept out of the equation. Others object that moral and political values have a legitimate, maybe even necessary, say in how we conceptualise the social world. A prominent advocate of the later position is Sally Haslanger. Haslanger proposes doing social ontology in a normative (or, in her words, "ameliorative") manner. According to this idea, we should define social concepts in a way that best suits our legitimate purposes, which may be a question of explicitly moral or political considerations.

Haslanger's proposal has been heavily criticised by Francesco Guala. According to Guala, the question how we should understand social kinds cannot be left "up to us" and the outcome of our potentially controversial moral and political disputes. Instead, he insists that social kinds, just like natural kinds, are determined by the nature of the external world, and our kind concepts ought to reflect that. Interestingly, both Haslanger and Guala rely on semantic externalism to defend their position. While Haslanger uses semantic externalist arguments to support the kind concepts she proposes, Guala argues that semantic externalism is incompatible with a normative approach to social ontology and points us towards his favoured form of realism about social kinds instead.

I point out that both Haslanger and Guala fail to appreciate the controversial status of semantic externalism even with regard to natural kinds. I argue that recent criticisms of semantic externalism lend support to the idea that moral or political values play an important role in scientific ontology. Values of this sort, so the idea goes, are needed to determine which set of properties associated with a paradigm sample is worth our scientific attention in the first place. In other words, moral or political values may play a legitimate *indirect* role in social ontology. A closer look at Haslanger's examples, however, suggests that she advocates using such values in a *direct* rather than indirect manner. These cases would require a different type of justification to the one provided by critics of semantic externalism.

In the following, I first introduce in more detail Haslanger's normative approach to social ontology. I then recapitulate Guala's objection to Haslanger's proposal, before

turning to recent criticisms of the semantic externalist position that influences both sides' arguments. I argue that these criticisms give traction to the idea that moral or political values play an indirect role in scientific ontology. I then point out that several of Haslanger's examples advocate a direct rather than indirect role of these values. Using the hybrid model of social kinds developed in Chapter 1, I clarify how exactly the cases differ, and explore how the direct influence of moral and political values could be justified.

3.1 HASLANGER'S ARGUMENT FOR VALUES IN SOCIAL ONTOLOGY

According to Haslanger, moral and political values should play an essential role in what social kinds we recognise in the first place. She proposes that we need to stop arguing about what certain kinds "really" are, and instead focus on what concepts we need to further our legitimate goals. For Haslanger, this means that moral and political considerations have a legitimate role to play in how we conceptualise the social world. Haslanger's examples range from race and gender to parent and marriage. Since race and gender are Haslanger's central case studies, we begin by focussing on these.

According to Haslanger, our legitimate purposes with regard to race and gender are to reveal and fight sexual and racial inequality. Terms such as "women" or "black" should be understood in such a way that best allows us to address these inequalities. These considerations motivate Haslanger's proposal of hierarchical race and gender concepts. The concepts define women, men and racialized groups in terms of being systematically subordinated or privileged on the basis of certain bodily features:[1]

> S *is a woman* iff S is systematically subordinated along some dimension (economic, political, legal, social, etc.), and S is "marked" as a target for this treatment by observed or imagined bodily features presumed to be evidence of a female's biological role in reproduction.
>
> S *is a man* iff S is systematically privileged along some dimension (economic, political, legal, social, etc.), and S is "marked" as a target for this treatment by observed or imagined bodily features presumed to be evidence of a male's biological role in reproduction.
>
> (Haslanger 2012, 230, original emphasis)

> A group is *racialized* iff its members are socially positioned as subordinate or privileged along some dimension (economic, political, legal, social, etc.), and the group is "marked" as a target for this treatment by observed or imagined bodily features presumed to be evidence of ancestral links to a certain geographical region.
>
> (Haslanger 2012, 236, original emphasis)

[1] Haslanger refines these concepts later on. For the sake of simplicity, I will stick to her preliminary definitions. "Iff" stands for "if and only if".

Haslanger acknowledges that race and gender terms may have different meanings in different contexts, and that the concepts above might not be suitable in certain contexts, such as the biology lab. Nonetheless, she suggests that these concepts have a claim to being the dominant understanding of race and gender both in the public domain and in the social sciences (Haslanger 2014, 111, 113). This implies that, on her view, our understanding of social kinds such as race or gender may reasonably depend on our moral and political values in social scientific contexts.

Interestingly, in addition to the normative argument above, Haslanger also provides a semantic externalist argument for her concepts. The semantic externalist argument suggests that there might be no need to rely on moral and political values when defending her concepts of race and gender. Instead, it states that these concepts make explicit what our race and gender terms have been referring to all along.

Before we consider this argument in more detail, it is useful to sharpen our understanding of semantic externalism. Semantic externalism states that the meaning of a term can be determined by facts "outside" the speaker, especially facts other than the speaker's beliefs about what the term means. It is a popular position in natural kind semantics. It has been argued that the meaning of natural kind terms is fixed not by speakers' intentions, but rather by the common "essence" of paradigm samples. Speakers point to these paradigm samples in an original "baptism event" (Putnam 1975, Kripke 1980). The original paradigm samples are often called "ostension". A common example for externalist natural kind semantics is the term "water". According to the semantic externalists, although competent speakers might simply think of water as watery stuff (colourless, odourless fluid that fills rivers and lakes, etc.), "water" really refers to H_2O. This, so the argument continues, is because H_2O is the common essence of objects in the ostension.

According to Haslanger, semantic externalism is not confined to natural kinds with internal essences[2], but can be applied whenever the ostension instantiates an "objective type". Objective types are groupings of entities that have a "degree of unity among its members beyond a random or gerrymandered set" (Haslanger 2012, 374). Haslanger suggests that gender and racial categories are not unified by biological essences, but still instantiate objective types. More precisely, she thinks individuals in these groupings share common features beyond the superficial morphological characteristics on the basis of

[2] There is an ongoing debate in what sense natural kinds have essences, and whether externalism in natural kind semantics requires internal essences (e.g. Boyd 1989; Häggqvist & Wikforss 2017).

which they have been classified as *women* or *white*. Categories like *women*, *black* or *white* are objective types in a social (rather than biological) sense. According to Haslanger, they are unified by the social characteristics of being subordinated or privileged along some dimension and by being targeted for such treatment based on one's (observed or imagined) physiological characteristics. In the case of the term "black" for instance, our linguistic intuitions might suggest that we refer to a biological kind. According to Haslanger, however, we are justified in saying that the term really is referring to is a social group that is systematically subordinated and marked for this treatment by physical characteristics that are taken as indicators of specific continental ancestry (Haslanger 2012, 236). Haslanger suggests that a similar argument can be made in the case of gender terms like "men" and "women" (Haslanger 2012, 230).

3.2 GUALA'S CRITICISM

According to Guala, there is an imminent tension between Haslanger's commitment to doing social ontology in a normative manner on the one hand, and her commitment to realism (as evidenced in the semantic externalist argument) on the other. He argues that, as social scientists, we want our classifications to reflect the underlying structure and mechanisms that sustain the phenomena we are investigating. The nature of a social kind, he suggests, cannot be at the same time determined by our norms and values and by the way the world is (Guala 2016, 190). Hence, the question what social kinds are is not "up to us" and our values. Instead, it is "constrained by properties and mechanisms that are not fully under our control, because they are determined by the external world" (Guala 2016, 190). Regarding the example of marriage, for instance, Guala claims that for a realist, "the meaning of marriage is determined by the way the world is, and the language that we use must be true to the way in which the (social) world is organized" (Guala 2016, 184-5).

Note that, in principle, Guala's objection is not limited to moral and political values. His assumption that the nature of social kinds is determined solely by the social world could equally be used to exclude the influence of so-called epistemic values. In the debate on values in science, epistemic values, such as accuracy, simplicity or explanatory, are often distinguished from moral or political values, the so-called "non-epistemic values" (McMullin 1982; Longino 1990). The idea behind this distinction is that, unlike non-

epistemic values, epistemic values are conducive to better scientific understanding.[3] For that reason, the influence of epistemic values on scientific reasoning and scientific ontology is much more widely accepted than the influence of non-epistemic values.[4] Since the influence of moral and political values is more controversial, and since Haslanger's and Guala's arguments are mainly concerned with them, the focus of this chapter will be on the role of non-epistemic values in social ontology.

Guala points out that saying that non-epistemic values should not influence social kinds is not to say that non-epistemic values have no justified influence at all. Instead, he suggests that normative and realist concerns each play a legitimate role in a different context. There is space for moral and political discussions about which specific form of a social kind we want to adopt in society. For instance, we should consider moral and political values when deciding which sorts of marriage we want to codify in the law. Nevertheless, Guala insists that questions about the nature of social kinds – such as "what is marriage?" – ought to be left to social scientists to be answered purely according to their best theories and evidence.

3.2.1 A tension between normative and realist concerns?

To appreciate Guala's critical reply to Haslanger, it is important to clarify how Haslanger understands the relationship between her normative argument and her semantic externalist argument. When reading Haslanger's case studies on race and gender, it seems that the normative argument and the semantic externalist argument are not in tension at all. But this is not for the reason Guala suggests, that is, not because Haslanger suggests using normative and semantic externalist considerations in different contexts. Haslanger does not believe that the semantic externalist argument is about the ontology of race and gender while the normative argument is about which concept of race and gender to adopt in our social practices.[5] Instead, Haslanger suggests that there is reason to believe that normative and semantic externalist arguments about race and gender point us in the same direction. Hierarchical concepts of race and gender are not merely the concepts that we should be using to identify and address sexist and racist oppression, they may also be the concepts that a semantic externalist analysis tells us we have been using all along.

---

[3] Note that some philosophers suggest rejecting the distinction between epistemic and non-epistemic values altogether (Rooney 1992). Nevertheless, the distinction is still commonly used today.

[4] For an "arch-realist" position that seems to reject any kind of value-influence, see Lewis (1999).

[5] To be clear, Haslanger does believe that these two questions can and should be distinguished. But making this distinction is not what resolves the potential tension between normative and semantic externalist considerations in her proposal for race and gender concepts.

For Haslanger, this is no coincidence. She argues that it may be the "job" of racist and sexist ideology to mislead us by suggesting that categories can simply be found in nature when we have actually played an important role in creating them (Haslanger 2012, 366, 383). In such cases, the primary normative demand on our conceptual apparatus may be to undo this masquerade and bring to light what our terms are actually referring to. This, she seems to suggest, is why the normative and the semantic externalist argument are likely to provide the same result in cases like race and gender.

Haslanger's approach to solving the potential tension between normative and semantic externalist considerations is not going to appease Guala. Although Haslanger suggests that normative and semantic externalist considerations coincide in the examples of race and gender, she offers no reason to think that the two can never come apart. Hence, even if we accept Haslanger's reasoning up to this point, there remains an obvious question: what should we do if the normative and the semantic externalist arguments point us to different directions?

For Haslanger, normative considerations have priority over realist ones. While she suggests that the prospects of making a semantic externalist argument in favour of her gender and race concepts are promising, she does not attempt to make these arguments in detail herself. Instead, she states that the primary aim of introducing the semantic externalist argument is to tackle a common objection to her race and gender concepts – the objection that these concepts are implausible because they run counter to our intuitions. According to Haslanger, the semantic externalist argument demonstrates that counter-intuitiveness is no reason for rejecting her concepts. Although the concepts are revisionary for conflicting with how ordinary speakers understand terms like "women" and "black", they may be non-revisionary in the sense that they make explicit what social kinds our race and gender terms have been tracking all along. In other words, Haslanger is using semantic externalist considerations mainly opportunistically to further her proposal for normative social ontology.[6]

Guala, by contrast, clearly prioritises realist concerns. He fears that letting non-epistemic values determine our understanding of the social world puts questions of social ontology at the mercy of potentially controversial moral and political debates. Guala suggests that the debate on marriage is a case in point. Gay rights activists and anti-gay

[6] Note that this is primarily the case for Haslanger earlier arguments (for instance Haslanger 2012, Chpt. 7). In later work, Haslanger is much more adamant that her race and gender concepts successfully explicate the meaning of race and gender terms on a semantic externalist account (Haslanger 2012, Chpt. 16). Here, it is no longer quite so clear whether Haslanger's thinks normative concerns should have the final say in social ontology.

religious conservatives both put forward definitions of marriage which are based on their conflicting moral and political values. If we commit to normative social ontology, it looks like we cannot determine what marriage is without first settling a normative dispute about the moral status of homosexuality.

According to Guala, this is an unhappy situation for social scientists – there are no signs that the normative dispute will resolve any time soon, and social scientists are neither experts nor authorities on questions of morality. Only a realist social ontology can save us from these troubles. We want social scientists, on the basis of their best theory and evidence, to give a single best answer to questions like "what is marriage?" just like natural scientists are able to provide a single best answer to the question "what is water?" (Guala 2016, 200).[7] For that, Guala insists, we need to put normative considerations to one side and focus exclusively on the semantic externalist project of finding out which "real" kinds our social kind terms are referring to (Guala 2016, 192). Doing social ontology in this sense, he suggests, "saves the realist principle that what marriage is, is primarily a scientific issue", and "opens the space for a descriptivist, scientific approach to the issue of same-sex marriage" (Guala 2016, 199). In other words, Guala argues that pursuing a strictly realist social ontology which is independent of moral controversies would allow us to determine the nature of marriage independently of any moral or political commitments.

The discussion suggests that there is an important connection between endorsing semantic externalism about social kind terms on the one hand, and arguing that non-epistemic values have no place in defining social kinds on the other. Guala believes that fully embracing a realist, semantic externalist position on social ontology means embracing the idea that the meaning of social kind terms is fixed unambiguously. In Guala's view, to say that there is only one correct or best realist understanding of social kind terms thus creates a stable bulwark against the influence of social or political values. But the argument has a weak point. While discussing the prospects and implications of extending semantic externalism from the realm of the "natural" to the realm of the "social", Haslanger and Guala both have lost sight of a crucial fact: even within its traditional remit of the "natural", semantic externalism is far from uncontested. This point is all the more important because the discussion exemplifies a tension that arises not only in the social sciences more generally, but also in other disciplines, such as psychiatry (see, for instance, Geuss 1981; Cooper 2007). In the following, I argue that recent criticism of

[7] Or at least scientists should be able to provide a single "best approximation that is available at a given time" (Guala 2016, 178).

semantic externalism about natural kind terms undermines Guala's argument for value-free social ontology.[8]

3.2.2 Criticism of semantic externalism about natural kind terms

To understand recent criticism of semantic externalism, it is useful to distinguish between a semantic and an ontological claim. The semantic claim states that the reference of a natural kind term is fixed by factors which are external to the minds of speakers, namely by the underlying "nature" or "essence" of the objects that the term has been paradigmatically applied to. The ontological claim, which is not always made explicit, states that the nature of this underlying "nature" is determined unambiguously by reference to a paradigm sample. Sören Häggqvist and Asa Wikforss argue that spelling out this ontological claim is important. It allows us to recognise that the success of semantic externalism heavily depends on a monist assumption about the ontology of natural kinds (Häggqvist & Wikforss 2017).

According to Häggqvist and Wikforss, the "guiding idea" of semantic externalism about natural kinds is that "once a sample (or set of samples) is fixed, the world accomplishes all the carving needed to determine extensions irrespective of human interests" (Häggqvist & Wikforss 2017, 17). In other words, the semantic externalist thinks there is one single correct or privileged way of distinguishing the objects in the world into water and non-water, marriage and non-marriage, etc. Furthermore, the semantic externalist assumes that this distinction is determined by a specific defining property that the objects in the paradigm sample share.

The first major objection to the premise that ostension to paradigm samples determines the referent of natural kind terms is the so-called "qua problem" (Devitt & Sterelny 1987). The qua problem states that ostension to a paradigm sample fails to fix the referent of a natural kind term, because the individuals in the sample will typically instantiate a number of kinds. For example, if we try to establish the kind term "giraffe" by simply pointing to a group of giraffes, the term could pick out any of a number of kinds or types instantiated by those individuals, such as *animal*, *mammal*, *herbivore*, *spotted thing*, etc.

The qua problem has pushed many people away from a purely externalist account toward a hybrid theory about the semantics of natural kind terms (e.g. Devitt and Sterelny

[8] There are further problems with Guala's proposal (see, for instance, Clarke 2017; Epstein 2016). These problems are not addressed here because they concern Guala's specific proposal, rather than his general defence of doing social ontology in a "strictly realist" manner. It is possible to endorse Guala's defence of a value-free realist understanding of social ontology without committing to his specific game theoretic account of social kinds.

1999; Stanford and Kitcher 2000).[9] On such a hybrid theory, the reference of natural kind terms is not solely determined by the nature of the paradigm samples, but partly depends on what descriptions speakers who baptise natural kind terms associate with these terms. The descriptions need to be specific enough to disambiguate the reference of the kind term. At the same time, they need to be general enough to allow for the possibility that speakers can be mistaken about the nature of the kind. After all, even proponents of hybrid theories want to be able to say that scientists *discovered* the nature of the stuff we have been referring to as "water", not that they have simply changed the meaning of our term. A possible candidate for a speaker description striking this balance is the idea of a causally specified placeholder (Devitt & Sterelny 1999, Stanford & Kitcher 2000). Roughly speaking, this view suggests that natural kind terms were intended to refer to whatever underlying physical or structural property is causally responsible for the observed properties of the individuals in the paradigm sample.

Without going into the details of specific causal placeholder proposals, we can note that there is reason to doubt that this response is available to Guala. The reason for this is that many kinds that play a central explanatory role in the special sciences cannot be identified with an underlying physical or structural property that all and only members of the kind possess (Dupre 1981, 1993; Millikan 1999, Boyd 1991). Instead, it has been argued that such kinds are better understood as homeostatic property clusters (HPC) (Boyd 1989, 1991, 1999). HPCs are constituted by a cluster of properties that are reliably but not perfectly co-instantiated, together with the causal factors or mechanisms which make it the case that the properties are reliably co-instantiated. As a result, HPC kinds generally cannot be identified in terms of a single underlying property. For HPC kinds, there is no straightforward way of telling how ostension to a paradigm sample could fix a term's meaning on a causal placeholder account.

Guala believes that social kinds are best understood as HPC kinds. On his realist account, social kinds are individuated in terms of their function, where the function consists in solving coordination problems. Guala argues that marriage, for example, typically solves several different coordination problems at once, such as procreation, child rearing and socialising, economic cooperation, and emotional support (Guala 2016, 198). At the same time, he observes that no single subset of these problems can be necessary and sufficient for marriage, because historical and anthropological evidence shows that humans have used contractual relationships to solve a variety of different combinations

[9] Hybrid theories of natural kind semantics are not be confused with the hybrid model of social kinds.

of these coordination problems. Guala therefore concludes that marriage "does not have a single essential function" and is better understood in terms of a cluster of coordination problems (Guala 2016, 199). In other words, Guala assumes that the social sciences need an HPC account of social kinds.

In order to support his strictly realist account of social ontology, then, Guala has to provide an account of how reference-fixing works for HPC kinds. But it is difficult to see how this could be done. HPC kinds, as argued above, cannot generally be identified with a set of shared observable properties. Instead, the motivating idea behind the HPC account is that, on a superficial level, many kinds featuring in the special sciences are "unified" merely by a form of family-resemblance of their members. Now, if certain objects that we find in the world are merely characterized by family-resemblance, it would not be possible to pick out a kind by pointing to a sample of them. Any attempt at drawing a kind boundary across the patterns of gradual change would be a matter of convention rather than something that reflects the structure of the world. Proponents of HPC, however, believe that there is an important difference between the patterns that characterise HPC kinds and mere family-resemblance. The idea is that HPC kinds are constituted by individuals that resemble each other due to the same sort of underlying mechanisms. In other words, HPC kinds are not individuated by specific sets of properties, but by clusters of properties that are reliably associated due to specific causal mechanisms.

Note that this qualification does not give us, and is not supposed to give us, sharp kind boundaries. The clustering produced by the mechanisms is still imperfect, hence might result in individuals that cannot clearly be assigned to one kind or another. But this is not a problem for semantic externalism, which can allow for the fact that natural kind terms might be vague. What semantic externalism needs, however, is the ability to identify some kind of boundary (sharp or not) in a non-arbitrary place. Otherwise, the groupings picked out as HPC kinds would not be groupings that reflect actual structures in the world at all, but rather a matter of convention. If that were the case, speakers' attempts at fixing the reference of natural kind terms by pointing to samples of these groupings would obviously fail. The hope of proponents of the HPC account, then, is that the mechanisms and causal factors that bring about the reliable clustering of properties can do this work for us. These mechanisms, so the idea, tell us where in a world of gradual change we can find natural boundaries, and they fix the reference of natural kind terms by attaching the terms to

those "underlying" causal factors and mechanisms that produce the reliable clustering of properties.

Unfortunately, people have expressed serious doubts on whether underlying mechanisms can do what proponents of HPC accounts are asking of them (Craver 2009, Häggqvist & Wikforss 2017). According to Carl Craver, the questions which mechanisms define a kind, when two mechanisms are mechanisms of the same type, and where one particular mechanism ends and another begins all cannot be answered without reference to conventional factors, such as specific human interests and purposes (Craver 2009). In other words, the problem of identifying natural boundaries, which reference to homeostatic mechanisms was supposed to solve at the level of clustering properties, reoccurs at the level of homeostatic mechanisms. The upshot is that we fail to individuate HPC kinds on the basis of specific homeostatic mechanisms, just as we failed to individuate them on the basis of specific properties.

According to Häggqvist and Wikforss, Craver's point has crucial implications for natural kind semantics – it makes semantic externalism about HPC kinds untenable (Häggqvist & Wikforss 2017). They argue that Craver's objection implies that it is not possible to point to an underlying mechanism that causally unifies the objects in the paradigm sample. As a result, ostension to such a sample does not fix the reference of natural kind terms unambiguously. Instead, we need to rely on further speaker-internal considerations to determine the reference of natural kind terms, such as their "overall function […] in ordinary classification, as well as how kinds are used and talked about in the sciences" (Häggqvist & Wikforss 2017, 18). This suggests that, at least for HPC kinds, there are serious doubts about the adequacy of semantic externalism.

3.3 SCIENTIFIC ONTOLOGY WITHOUT SEMANTIC EXTERNALISM

It is time to recapitulate the main points from the discussion above. We established that Guala's "realist" objection to normative social ontology hinges on the semantic externalist assumption that there is one correct or privileged meaning of (social or natural) kind terms. This privileged meaning, so the semantic externalist story goes, has been established in a baptism event by ostension to paradigm samples. We then observed that establishing the meaning of kind terms cannot be quite so straightforward because individuals in the paradigm sample typically instantiate a variety of kinds. This made semantic externalists gravitate toward causal placeholder accounts. According to these

hybrid accounts of natural kind semantics, the meaning of a natural kind term is partly determined by speakers' intentions to pick out whatever causal features explain the shared observed properties of individuals in the paradigm sample.

Without going into the general problems associated with causal placeholder accounts, we noted that this strategy is unlikely to be available to someone who, like Guala, defends semantic externalism in the context of an HPC account of natural/social kinds. This is because the HPC account owes an explanation as to how the underlying mechanisms "carve nature at its joints". It is not clear how these mechanisms themselves can be typed and individuated without reference to conventional elements such as explanatory interests. Without such an explanation, semantic externalism about HPC kind terms cannot avoid collapsing into a conventionalist account, according to which the meaning and extension of kind terms is at least partly determined by our goals and interests.

Once Guala is forced to admit that the meaning of social kind terms inevitably depends on explanatory interests and purposes, it becomes much harder to insist that non-epistemic values must be kept out of the process. This is because non-epistemic values often legitimately shape our explanatory interests and purposes. I will say more on this below. Of course, Guala could still insist that the true meaning of social kind terms is fixed by whatever explanatory purposes or interests featured in the baptism event. There is merit in this position, as we will see further below. Once the meaning of a social (or natural) kind term has been determined in the light of specific explanatory aims or interests, there is good reason to argue that we cannot simply change the term's meaning in the light of different aims and interests. Yet, by admitting that the meaning of these terms are essentially a matter of which interests "were there first", this response fails to save Guala's realist commitment. The meaning of social kind terms can no longer be determined by the structure of the external world alone.

3.3.1 Avoiding "anything goes" in grouping and labelling

The discussion in the previous section suggests that there is a serious problem with Guala's argument against normative social ontology. Guala attempts to extend semantic externalism from natural kinds to social kinds without recognising the unresolved challenges semantic externalism faces even with respect to natural kinds. In response, I suggest that we reject Guala's argument and instead return to the worries that motivated it. Guala fears that, unless semantic externalism succeeds, we are left with an implausible "anything goes" account of scientific ontology:

> To be a realist is precisely to believe that reference is determined by the world, if anything, and it is not "up to us […]". According to the realist, […] the identity of any entity is robust to a limited range of manipulations only, so there are changes to [a social kind] that will necessarily turn it into something else. Just as the term "senator" cannot be stretched so far as to include Caligula's horse, so there are arrangements that cannot be legitimately called "marriage," regardless of what we want the latter to be.
>
> (Guala 2016, 190)

In other words, Guala worries that once we reject a semantic externalist account and allow that the meaning of social kind terms can be influenced by our values and interests, we throw out of the window all worldly constraints on how to classify social entities. As a result, we risk ending up with a social ontology that looks entirely unfit for scientific purposes.

This is quite a serious concern, but is it justified? To answer this question, it is useful to distinguish two worries in Guala's characterisation. The first worry is that there must be some worldly constraints on what sort of entities can sensibly be *grouped* together. Guala suggests that we should be suspicious of a position which – assuming our interests dictate it – would allow for the British Show Pony Society, the German Parliament, and Victoria sponge cake to be grouped together as a single social kind. Call this the *grouping worry*.

The second worry is slightly different. It concerns how far we can extend existing *labels* such as "marriage" to refer to different groupings, hence will be called the *labelling worry*. The labelling worry is independent of the grouping worry. For instance, even if it was legitimate to group together the British Show Pony Society, the German Parliament, and Victoria sponge cake, we might have good reasons for not wanting to call that grouping "marriage."

In light of this distinction, we can see that the motivating concern behind Guala's defence of a strictly realist position is to avoid collapsing into an "anything goes" position with respect to both grouping and labelling. In the next section, I argue that Haslanger recognises and attempts to address both the grouping and the labelling worry. Finding her responses insufficient, I discuss how they can be modified so as to meet Guala's concern.

3.3.2 Responding to the grouping worry

The grouping worry is addressed in Haslanger's objective type constraint on social kinds discussed above (Haslanger 2012, 200-10). The objective type constraint says that, although social kinds need not be defined in terms of an underlying "essence", they do have to track real distinctions in the world. In other words, the objective type constraint

demands that the entities that we group together in a social kind at the very least need to have certain properties in common.

At first glance, the objective type constraint might look like a promising way of addressing the grouping worry. Objective types, so it seems, cannot be constituted by arbitrary conglomerations of entities. But a closer look suggests that there is a serious mismatch between Haslanger's objective type constraint and the constraints that are commonly put on scientific ontology. The objective type constraint is at the same time weaker and stronger than widely accepted scientific constraints. Given that the motivation behind the grouping worry is the need to group entities in a way that is useful from a scientific point of view, this would mean that Haslanger's objective type constraint is inadequate.

Consider the problem in more detail. The objective type constraint is stronger than commonly accepted constraints on scientific ontology because it demands that there is a property that all and only entities in the social kind have in common. The constraint thereby falls behind the widely endorsed HPC account of scientific kinds. According to the HPC account, there need not be a single property that all and only members of the kind share. Instead, natural kinds are characterised by a cluster of reliably associated properties, none of which are individually necessary for kind membership. Moreover, proponents of the HPC account argue that HPC kinds are perfectly adequate for a number of scientific endeavours because they facilitate inductive inferences, generalisations, and predictions. By putting a constraint on social ontology that rules out HPC kinds, Haslanger's constraint is unnecessarily restrictive.

At the same time, the objective type constraint is weaker than widely accepted scientific constraints. This is because it allows for social kinds to be unified by a single, relational property. Social kinds which are unified only by a relational property could be highly arbitrary from a scientific point of view. Haslanger suggests that her objective type ontology would include sets such as "the set of objects currently on my kitchen counter" and "the set of numbers randomly selected by the New York State Lottery on January 1, 2011" (Haslanger 2012, 202, 208). Since these two examples are no less arbitrary than the set of The British Show Pony Society, the German Parliament and Victoria sponge cake (unified, say, by the relational properties of entities I am currently thinking about), the constraint clearly fails to address the grouping worry in any sufficient way. It is unable to tell us how we can balance the idea that social and natural kinds are inevitably grouped in light of our values and interests with the requirement that the resulting groupings be

suitable for scientific inquiry. In the following section, I develop a response to the grouping worry by distinguishing direct from indirect influence of non-epistemic values.

3.3.3 Direct and indirect influence of non-epistemic values

Consider the following general model for an indirect role of non-epistemic values. According to this model, we can understand the process of choosing scientific ontology as a two-stage process. At the first stage, we filter out all those groupings that are not scientific kinds. "Scientific kinds" here can be understood, as broadly as possible, as groupings that are suitable for epistemic purposes. The scientific kinds requirement is meant to reflect the commitment that the kinds that feature in the sciences group entities in such a way as to allow us understand, explain, predict, or control certain aspects of the world. There is substantial disagreement about what sort of groupings are best suited for these purposes (Boyd 1989; Khalidi 2013; Magnus 2014; Slater 2015; Ereshefsky & Reydon 2015), but the scientific kinds requirement can be agnostic about which of these accounts is most adequate (maybe they all are for different scientific purposes). The point it makes is that, in order to qualify as a scientific kind, entities need to be grouped in a way that facilitates epistemic purposes. The distinction which the scientific requirement makes might be vague, but it is specific enough to allow us to disqualify groupings that clearly fall on the wrong side. Hence, "the set of objects currently on my kitchen counter" and "the set of numbers randomly selected by the New York State Lottery on January 1, 2011" do not qualify as scientific kinds because such groupings are not fit for epistemic purpose. They will not allow us to understand, explain, predict, or control certain aspects of the world.

The problem with the scientific kinds requirement is that it cannot explain how we end up with a specific limited set of kinds that allow communication and progress in the sciences. There are potentially innumerable aspects of the world that we could attempt to understand, explain, predict or control. Accordingly, even after filtering out all non-scientific kinds, chances are that we will still be left with a huge number of scientifically adequate groupings with respect to any previously established putative kind. Luckily, there is a widely accepted response to this problem. At least among philosophers of the special sciences, it is fairly common to think that the question which scientific groupings we ultimately adopt as social or natural kinds is a matter of our explanatory interests (e.g. Kitcher 1984, Brigandt 2009). In other words, scientific kinds which we actually use in the special sciences need to be constructed in such a way as to allow us to understand, explain, predict, or control aspects of the world which are of interest to us.

This is the second stage of the model, and the stage where non-epistemic values come into play. Explanatory interests, so the idea, do not simply fall from the sky. As demonstrated in various case studies, explanatory interests are often, and possibly inevitably, influenced by non-epistemic values (Anderson 1995, 2004; Longino 1990, 2013; Gannett 2010; Kitcher 2011; Brigandt 2009; Ludwig 2017). In other words, the model suggests that non-epistemic values play a legitimate, maybe even necessary role in scientific ontology, because they decide which aspects of the world we care to investigate in the first place. This means that the role of non-epistemic values is constrained in a crucial manner. Non-epistemic values may only have an *indirect* influence on scientific ontology. Non-epistemic values decide which epistemic goals and interests determine our scientific ontology, but they do not shape scientific ontology directly.

Incidentally, what it means for non-epistemic values to exert indirect influence on social ontology is well illustrated by Haslanger's central examples of race and gender. Recall that, according to Haslanger, race and gender are not defined by biological characteristics such as morphology or genes, although these characteristics are relevant in a derivative sense. Haslanger argues that race and gender should be defined in terms of social positions of systematic privilege/subordination that individuals are allocated to on the basis of certain (real or imagined) bodily features. Moreover, she suggests that grouping individuals in this way is useful to advance specific epistemic aims. For Haslanger, this is first and foremost the aim of identifying and explaining racial and sexual inequalities, which includes elucidating "how social forces, often under the guise of biological forces, work to perpetuate such inequalities (Haslanger 2012, 226-7).

Note that Haslanger's reasoning is not trivial or circular because it does not presuppose the race and gender kinds just defined. "Racial and sexual inequality" is not to be understood as inequality between her race and gender kind that are, per definition, systematically privileged or subordinated. In the case of gender, for examples, Haslanger clarifies:

> Given the priority I place on concerns with justice and sexual inequality, I take the primary motivation for distinguishing sex from gender to arise in the recognition that males and females do not only differ physically, but also systematically differ in their social positions. What is of concern, to put it simply, is that societies, on the whole, privilege individuals with male bodies.
>
> (Haslanger 2012, 229)

In other words, the sort of inequality that Haslanger's concepts aim to identify and explain is best understood as a systematic inequality between individuals of different types of bodies.

Importantly, although these definitions of race and gender are obviously influenced by non-epistemic values, such as the commitment to social equality, the role played by these values is merely indirect. The commitment to social equality motivates the epistemic aim of identifying and explaining racial and sexual inequality. Haslanger's definitions of race and gender are designed to further these epistemic aims. But the value of social equality does not directly influence who gets to be classified, say, as a *woman*.

To make the contrast clear, it is useful to consider what it means for non-epistemic values to directly influence ontological choice. Non-epistemic values have a direct influence on our ontological choices if they do not take a detour via explanatory aims. This means that we group entities in a certain way not because it allows us to understand, explain, predict, or control aspects of the world that will help us realise moral or political values, but because it directly furthers moral or political purposes. An example of this would be the decision to classify the Alabama Sturgeon as a species because it would allow us to implement conservation measures reserved for endangered species (see Scharpf 2000, quoted in Ludwig 2017, 1263).

While many people accept the indirect influence of non-epistemic values on ontological decisions as an epistemically legitimate and probably even necessary aspect of scientific inquiry, the direct influence is much more controversial.[10] It is easy to see why. Individuals grouped together according to non-epistemic criteria will not generally share the properties that are epistemically relevant to the scientific project at hand. If social and natural kinds have been carefully grouped so as to facilitate specific epistemic aims, any tampering with these groupings under the direct influence of non-epistemic values will likely render them less suitable for furthering these epistemic aims than they were before.

In sum, the discussion above suggests we can recognise a legitimate role of non-epistemic values in social ontology while answering Guala's grouping worry. The solution is to limit the influence of non-epistemic values to an indirect role. Objects are grouped into natural and social kinds in light of what we are interested in understanding, explaining, predicting or controlling, but what we are interested in understanding,

[10] David Ludwig (2017) argues that we need to accept both the direct and indirect influence of non-epistemic values on ontological choices. However, his argument only shows that direct influence does happen, not that it is a desirable or necessary element of scientific ontology.

explaining, predicting or controlling is often informed by our moral and political values. Since social and natural kinds will only be able to facilitate these purposes if they group entities with respect to relevant properties and relations, arbitrary conglomerations as the ones described above are excluded from scientific ontology.

3.3.4 Addressing the labelling worry

To fully answer Guala's concerns, we still need to address the labelling worry. This means we need to know under which circumstances we are can use existing terminology to refer to groupings other than those the terms traditionally refer to, and when we should be using new terms instead. We can address this problem by returning to Haslanger's race and gender concepts. As noted at the beginning of this chapter, Haslanger not merely proposes a new way of grouping individuals in the light of normative interests. She also suggests using existing terminology for the job, namely familiar terms like "woman", "man", "black", etc. As Haslanger acknowledges, demonstrating that a novel way of grouping individuals supports a specific epistemic aim is not the same as demonstrating that we are justified in appropriating existing terminology to refer to the new grouping (Haslanger 2012, 225). How exactly does Haslanger justify using terms like "woman" and "black" for her novel groupings, especially considering the looming confusion with ordinary understandings of these terms? Would we not all be better off introducing new labels for this purpose?

According to Haslanger, the answer to that question depends on two conditions, a political and a semantic one. The political condition involves considerations such as "the acceptability of the goals being served, the intended and unintended effects of the change, the politics of the speech context, and whether the underlying values are justified" (Haslanger 2012, ibid.). The semantic condition requires that "central functions" of the term are preserved. Haslanger suggests that preservation of central functions can be achieved, for instance, if the terms continue to be used to "organize or explain a core set of phenomena that the ordinary terms are used to identify or describe" (Haslanger 2012, 225).

Further detail would be needed to decide whether these conditions can disperse Guala's concerns. Unfortunately, Haslanger does not discuss them in any more depth. With regard to the political condition, this may be understandable. The political condition involves criteria that are highly context-dependent, such as unintended effects and politics of the speech context. For that reason, it may not be possible to give a more detailed, general description. Instead, the political condition will likely have to be assessed in a case-

by-case manner, consulting available empirical evidence.[11] This is not the case, however, for the semantic condition. Here, it should be possible to elaborate on a general theoretical level what it means for a term to continue to organise or explain the same "core set of phenomena" that it describes in ordinary language.

One interpretation that immediately comes to mind is that the semantic condition is in fact a semantic externalist condition. On this view, the semantic condition states that terms can be appropriated if they continue to refer to the same paradigm sample "essence". This would be a bad move for several reasons. For one thing, it would undermine the coherence of Haslanger's argument for doing social ontology in a normative as opposed to realist/semantic externalist manner. By reintroducing semantic externalism through the back door, Haslanger would admit that normative social ontology is eventually bound by the constraints of semantic externalism.

More importantly, interpreting the semantic condition as a semantic externalist condition would mean that Haslanger's normative social ontology is subject to the same problems as semantic externalism. Recall that the core problem for semantic externalism about kind terms is that pointing to a paradigm sample alone leaves the meaning of the term in question underdetermined. If Haslanger's semantic condition is understood as a semantic externalist condition, the only constraint it would put on terminological appropriation is that new definitions of a term must include individuals from the paradigm sample.

To see this, consider the case of "woman". Assume that the paradigm sample in the baptism event for the term "woman" consists of a group of human individuals with female anatomy. In light of the problems with semantic externalism, this would mean that we could use the term "woman" to refer to such diverse groupings as mammals, humans, humans with female anatomy, humans between the age of 25 and 35 years, residents of the country that the individuals pointed to happen to live in, and so forth. In other words, if Haslanger's semantic condition is interpreted as a semantic externalist condition, it would be far too permissive to do the required work. But all is not lost. In the following, I use the lessons from the criticism of semantic externalism to outline how the semantic condition can be made to work.

[11] See Saul (2006) for a more detailed discussion of this point.

3.3.5 Explanatory interests and semantic variation

The discussion of semantic externalism showed that ostension to a paradigm sample determines the meaning of a term only in the light of epistemic aims or interests that decide which of the properties shared by the paradigm sample are relevant. This fact can be used to make sense of Haslanger's semantic condition for appropriating existing terminology. According to the semantic condition, appropriating a kind term is justified if central functions of the term are preserved. As argued above, it is only with the help of epistemic interests that ostension to a paradigm sample can determine the meaning of a kind term. To make Haslanger's semantic condition work, then, the key is to understand the "central functions" of a term in terms of the epistemic aim or interest that it serves. Accordingly, we should modify the semantic condition as follows: it is permissible to alter the definitions or concepts associated with existing scientific terminology as long as the new definitions serve a term's existing epistemic aim.[12]

The point can be illustrated using the familiar examples of race and gender. Plausibly, the epistemic aim – or at least one of the epistemic aims – of using terms like "men", "women", "black" and "white" is to explain patterns of human variation. In light of this epistemic aim, the question is what concepts or definitions of "women", "white", and so forth scientists should be using in order to understand and explain patterns of human variation. Traditionally, scientists approached the task by using biological definitions. They defined "men" and "women" in terms of biological reproduction, and racial terms in terms of continental ancestry. The reason for this was that they thought that underlying biological characteristics such as genes and hormones are causally responsible for the (assumed or observed) differences in behaviour and psychology.

But the success of these definitions in furthering the epistemic aim of understanding patterns of human variation has been somewhat mixed. Generally speaking, biological definitions of sexual categories have turned out to be much more successful than biological definition of racial categories. The former allowed scientists to make numerous inferences and explanations about physiological differences and are well-established categories in, say, medicine. The latter had a much more limited inferential and explanatory use, and the few cases where scientists claimed epistemic success tend to be highly contested (Gannett 2005; Bolnick 2008; Feldman & Lewontin 2008). At the same time, it has become apparent that for many observed differences between members of

[12] For a more developed version of this outline proposal, using the example of the molecular gene concept, see Brigandt 2010.

racial and sexual categories, explanations in terms of underlying biological categories are either misguided or highly incomplete. Instead, many race or gender differences of behaviour, psychology and even health may be a function of differential socio-economic status, prejudice and discrimination, rather than biology (Kaplan 2010).

On the basis of these observations, Haslanger's appropriation of race and gender terms could be justified as follows: The epistemic aim of race and gender terms is to understand and explain variation between racial and sexual groups (originally identified in terms of morphology). There is reason to believe that a number of differences between these groups are best explained by reference to social factors, in particular those associated with systematic privilege or disadvantage. Accordingly, biological definitions of terms like "women" or "black" will be of limited use in achieving the term's epistemic aim. They will not help us identify and understand the social determinants of variation between racial and sexual groups. To elucidate these aspects of human variation, we should define categories like "men" or "black" in terms of the systematic privilege or subordination that individuals acquire on the basis of morphological markers.

In the above argument, the semantic condition for terminological appropriation is fulfilled because the terms in question continue to serve the same epistemic aim. Terms like "women" and "men" have always been used to explain patterns of difference and inequality, Haslanger's redefinition merely redirects our attention to a different set of potential causes. The discussion shows that we can interpret Haslanger's semantic condition in a way that answers the labelling worry. The response will appease the realist, because the influence of non-epistemic values on labelling decisions is only an indirect one. It allows us to alter the definitions of existing scientific terms if it promotes the epistemic aim already associated with those terms. But it leaves no room for non-epistemic values to influence our labelling choices directly.

3.4 DIRECT INFLUENCE OF NON-EPISTEMIC VALUES REVISITED

3.4.1 Reconsidering Haslanger's examples

It is time to briefly recapitulate the arguments in the previous sections. Having rejected semantic externalism and distinguished direct from indirect influence of non-epistemic values, the case for the influence of moral and political values on social ontology became harder to reject and, at the same time, less threatening for proponents of a realist social ontology. I illustrated this by distinguishing between a grouping and a labelling worry

behind Guala's concern about normative ontology. The grouping worry, which is about individuals being grouped in a way that is arbitrary from a scientific point of view, can be answered by limiting the influence of non-epistemic values to an indirect role. Kinds need to be grouped in the light of epistemic interests, but these interests can in turn be motivated by non-epistemic values. The labelling worry, by contrast, expresses a concern about stretching the definitions of existing scientific terminology beyond recognition. It can be answered by demanding that changes in the concept associated with a term need to be made in a way that contributes towards the term's existing epistemic aim. Since both responses allow non-epistemic values to influence grouping and labelling scientific kinds only in an indirect manner, they should be acceptable to a proponent of realist social ontology.

This suggests that the only potential point of conflict that remains between realist and normative approaches to social ontology concerns the direct influence of non-epistemic values on ontological choices. While realists think that non-epistemic values may influence grouping and labelling choices only indirectly via epistemic aims, proponents of normative ontology might be inclined to say that direct influence is permissible, too. The discussion up to this point, however, could suggest that there is no such conflict between Guala and Haslanger. Instead, it seems that Haslanger, who is the main subject of Guala's criticism, recommends only the indirect influence of non-epistemic values. Her main examples of race and gender consistently served us as case studies illustrating indirect influence of non-epistemic values. In other words, it seems that Haslanger's social ontology is normative only to an extent that realist like Guala cannot really take issue with. After all, Haslanger seems to merely advocate the indirect influence of political or moral values in determining what epistemic purposes our ontological choices serve in the first place. This impression, however, is misleading. Looking beyond the examples of race and gender, we can see that not all of Haslanger's case studies follow the pattern described above.

To see this, return the example of marriage discussed at the very beginning of the chapter. On the realist view, it would not be permissible to include same-sex relationships in the social kind *marriage* simply for the immediate moral and political aim of furthering social equality. Doing so would mean letting one's moral or political values directly influence ontological decision. But this seems to be exactly what Haslanger suggests. She argues that questions such as whether marriage happens only between a man and a woman "are not plausibly semantic controversies, but are social and political ones" (Haslanger 2012, 433). In other words, Haslanger believes that the question whether same-sex

relationships are marriages is not a matter of facts (be they semantic or scientific ones), but of moral and political values. It is not a question about how the world is, but about how we want it to be.

Are Haslanger's claims about the normativity of the kind *marriage* in conflict with the realist commitment? Perhaps not inevitably. One simple way to avoid conflict might be to extend the range of purposes our ontological choices may serve. Once we acknowledge that social kinds can be understood differently in different scientific contexts, we may only be a small step away from saying that social kinds might be understood differently again in non-scientific contexts. The realist commitment is a commitment about scientific ontology, but we do not always group entities for scientific purposes. In non-scientific contexts, the primary purpose in using social kinds may not be epistemic. Instead, we might put them together for explicitly political and pragmatic ends, such as organising social life.

According to this view, it would be misguided to apply the same demands on social ontology in both scientific and non-scientific contexts. While epistemic concerns may have priority in scientific contexts, where the direct influence of non-epistemic values on ontological choice will generally be detrimental to our aims, in certain public contexts the direct influence of moral or political values may be exactly what we need. In the case of marriage, an argument along these lines would go something like this: (i) in the public realm, we need to make ontological choices that further social equality; (ii) including same-sex relationships in the kind marriage would further social equality; therefore, (iii) same-sex relationships should be included in the kind marriage. But this argument is vulnerable to criticism. A proponent of a conservative account of marriage could argue that social equality merely demands that the legal status associated with marriage becomes available to same-sex couples. This purpose, the conservative could argue, is achieved once same-sex couples are able to enter civil partnerships which entail the same rights and duties as marriage. It does not require to extend the term "marriage" to include legal partnerships between same-sex couples. In the remainder of this chapter, I use the hybrid kind model to show why this response is unconvincing.

3.4.2 Hybrid kinds and the direct role of non-epistemic values

To illustrate the point, it is helpful to briefly move from marriage to an example which Haslanger discusses in more detail – the example of the kind *parent* (Haslanger 2012, Chpt. 14). Haslanger observes that primary school notes are usually addressed to "parents", for instance in the form of invitations for "Parent Nights", "Parent Breakfasts", or "Parent-

Teacher-Conferences". According to Haslanger, there is a mismatch here between the common understanding of the term "parent" on the one hand, and the use of the term in school notes on the other. If we asked people what "parent" means, they would give an answer along the lines of biological parent, i.e. immediate progenitor. But this seems to be at odds with how the term operates in the context of the primary school notes. Here, we would expect that "parent" refers to primary caregivers, no matter whether they are biological parents, step parents, legal guardians, and so forth.

Haslanger suggests that there are several possible ways of responding to this mismatch: (i) the school could insists that "parent" in their notes refers to the common understanding of biological parent; (ii) the school could address the notes to "Primary Caregiver" instead; or (iii) we could collectively alter our understanding of "parent" so as to refer to primary caregivers instead, at least in public contexts such as primary schools. While Haslanger is quick to dismiss the first option as "clearly misguided […] as a social/political matter", she suggests that deciding between options (ii) and (iii) is somewhat trickier (Haslanger 2012, 390).

For the sake of clarity, it is worth discussing the three options in more detail. Based on what we have established so far, there are two different grounds for rejecting option (i). We could imagine that the primary school really intends to only invite biological parents to the events. In that case, option (i) would be "fit for purpose" but we would be inclined to criticise the purpose at hand on moral or political grounds. For instance, we could argue that inviting only biological parents to the primary school events is harmful and discriminating towards children who do not grow up in traditional family arrangements, as well as towards their "non-traditional" primary caregivers. This is what Haslanger seems to have in mind when she rejects option (i). But there is a second, somewhat more charitable interpretation. Assume the primary school really wants to facilitate socialising and communication with whoever has primary responsibility in looking after the children. In that case, we should reject option (i) not on moral or political grounds, but for reasons of coherence. The reason the school cannot insist that "parent" refers only to immediate biological progenitor is because doing so would conflict with what they aim to communicate. In other words, the school cannot at the same time insist that "parent" refers to biological progenitors and continue using the terminology in its notes, because doing so would be inconsistent with their pragmatic aim in writing the notes.

The example is instructive because it suggests that the main point of conflict between proponents of normative and realist social ontology concerns labelling rather than

grouping. The idea that we can group individuals in the light of non-epistemic aims – such as the aim to facilitate communication with whoever has primary responsibility in looking after children in a primary school – is quite uncontroversial. It is only when we try to label those groupings with existing terminology that we enter contested territory. This is where options (ii) and (iii) differ. Both acknowledge that the grouping commonly referred to as primary caregivers is the right one for the job at hand, but they disagree as to whether the term "parent" should be used to refer to that grouping. Haslanger states that it is not clear from her example which is the best strategy to adopt overall. Her point is merely to demonstrate, against the terminological conservative, that there may be considerations which point towards option (iii), i.e. maintaining the original term and modifying the associated meaning instead. Although Haslanger acknowledges that the decision to redefine terms must be made on a case-by-case basis, she believes there can be considerations that speak strongly in favour of doing so.

In order to understand what exactly these considerations are, and how they relate to the ontology of the kind in question, we need to dig a little deeper. In the case of parent, Haslanger argues that "the term 'parent' brings with it a certain normative weight, entitlement, and so on, that the term 'primary caregiver' doesn't" (Haslanger 2012, 401). Similar remarks lead us back to our original example of marriage. According to Haslanger, there is a reason why gay activists are fighting over the meaning of "marriage", rather than simply opting for a different term like "civil union". The term "marriage", she argues, "links the institution to a broad range of other social phenomena, and does so in a way that 'civil union' cannot approximate" (Haslanger 2012, 402). In other words, Haslanger suggests that the decision between using new or existing terminology is not simply about words, but may have a wide-ranging impact on people's lives.

The hybrid model allows us to understand how a seemingly simple terminological choice can bring about such an impact. As shown in Chapter 1, the model describes how a specific grouping (the base kind) can be associated with a specific status (the status kind). The status kind consists of a cluster of norms, expectations, entitlements, etc. are socially associated with the grouping at hand. *Parent* and *marriage* can be understood in terms of these hybrid kinds. Presumably, the terms "parent" and "marriage" were originally used simply to denote specific groupings of entities – biological progenitors in the former case, and something like a formally recognised economic, sexual, and emotional relationship between a man and a woman in the latter case. Yet, over time these classifications developed important lives of their own. Through being classified as a parent, or as

married, individuals simultaneously acquired membership in a status kind that locates them in a complex network of social relations, expectations, informal obligations, entitlements, and so forth. As argued before, this not only involves how one is viewed by others (and oneself) but also regulates access to various social resources and opportunities.

Understanding *parent* and *marriage* as hybrid kinds allows us to see how non-epistemic values can be directly relevant to our terminological choices. Denying, say, adoptive parents the title "parent", or gay couples' legal relationships the title "marriage" also means denying them the social status that is associated with these labels. Since denying individuals the social status associated with such key social terms may have a far-reaching impact on their lives, the terminological choices in question can be of significant moral and political import. This argument provides a strong case for the position that moral and political considerations are *directly* relevant to terminological choice.

An opponent of normative social ontology could respond that these concerns may legitimately influence terminological choices in the public realm, but they should have no impact on scientific terminology. They could argue that, in public contexts, where the benefits or disadvantages associated with being labelled in a certain way accrue, our use of terms like "parent", "marriage" may directly depend on moral or political considerations. In scientific contexts, however, terminological choices should only depend on our epistemic aims. This response, however, presupposes a clear separation of public and scientific spheres that might be unrealistic. It disregards a potential tendency of the public to turn to scientists as experts on what terms like "marriage", "parent" or "woman" mean. At the same time, it ignores the potential impact public labelling practices may have on the features of individuals studied by the sciences.[13]

I will not attempt to draft general guidelines as to when the direct influence of such values is justified, that is, if and under what circumstances the described impact of classifications onto individuals should lead us to modify classificatory practices in scientific contexts. As mentioned above, the answer is likely to be quite context-dependent and would lead us too far into the realm of moral and political philosophy.[14] Instead, I will instead take a closer look at the epistemic challenges that arise from this hybrid structure in the next chapter. What I hope to have demonstrated in this section is twofold. First, whatever reason we might have to preserve existing practices of labelling,

[13] I will examine these interactions in greater detail in the next chapter.
[14] For an insightful discussion of the moral and political concerns associated with the status of *woman*, see Jenkins (2016).

there may be substantial moral and political considerations pointing the other way. Second, the hybrid kind model provides a solid ontological footing for understanding and addressing these complications.

3.5 CONCLUSION

In this chapter, I used the hybrid kind model to illuminate the question whether there can be a "strictly realist" social ontology void of any influence of non-epistemic values. I started the inquiry by considering Haslanger's race and gender concepts, for which Haslanger offered both a normative and a semantic externalist argument. The normative argument suggests that we need to define social kind terms in the light of moral and political values, whereas the semantic externalist argument suggests that the meaning of social kind terms is determined by ostension to paradigm samples. Guala criticised that the normative and the semantic externalist are in tension. He proposed that this tension should be resolved by recognising that the meaning of social kind terms – just like natural kind terms – is determined purely by semantic externalist considerations.

I argued that Guala's objection to normative social ontology is unconvincing because it fails to engage with existing criticism of externalist natural kind semantics. A more detailed look at this debate suggested that semantic externalism (and its causal placeholder variation) are untenable for the HPC account of kinds favoured by Guala. I then argued that Guala's realist commitment can be preserved in a weaker form, which requires that grouping and labelling choices in scientific contexts need to be made in the light of purely epistemic aims. Non-epistemic considerations, such as moral and political values, are allowed to influence ontological choices only indirectly, by determining which epistemic aims we decide to pursue in the first place.

It turned out, however, that this understanding still leaves room for conflict between Guala's and Haslanger's positions. Although this is not the case for the examples we started out with – Haslanger's race and gender concepts, so I argued, are prime examples of the indirect rather than direct influence of non-epistemic values – Haslanger is clearly suggesting a direct role for non-epistemic values in her discussion of "parent" and "marriage." Using the hybrid kind model developed in Chapter 1, I suggested that Haslanger's point is best understood in light of the social statuses associated with our labelling practices. The fact that our terminological decisions may have a wide-ranging impact on people's lives suggests that moral and political considerations may legitimately play a direct role in these decisions. In other words, there is reason to think that non-

epistemic values may influence our classificatory practices not only indirectly but also directly.

4

CAPRICIOUS KINDS

In order to understand the full scope of the epistemic challenges posed by hybrid kinds, it helps to approach the phenomenon from a slightly different angle.[1] While the previous chapters focussed mainly on social kinds, that is, kind classifications used in the social sciences, this chapter will take a step back and consider classifications in the human sciences more generally. "Human sciences", here, is an umbrella term for any scientific disciplines that studies humans. It therefore encompasses a wide range of subjects ranging from sociology, history and psychology to medicine and genetics. The classifications used in these sciences to classify human individuals, traits or behaviours will be referred to as "human kinds".

The question whether the human world can be studied in the same way as the non-human natural world has given rise to several heated controversies over the last two centuries. On the one side, proponents of the *unity thesis* argue that investigation of the human world ought to be modelled closely on our scientific methods for the investigation of the natural world. On the other side, proponents of the *difference thesis* defend the idea that the human world is importantly different from the natural world, and therefore requires methods fundamentally different from those of the natural sciences. Today, this highly polarised characterisation looks somewhat outdated. For better or worse, grand claims about the nature of "the natural" sciences as opposed to "the human" sciences have given way to a more nuanced investigation of specific scientific disciplines and approaches. Accordingly, the idea that the investigation of the human world requires a fundamentally different approach to that of the natural sciences has become a minority view in philosophy of science.

One of the last spokespersons of this view is Ian Hacking. For Hacking, the special status of the human sciences lies with the kinds they study: while the kinds that figure in the natural sciences are independent of (or, in Hacking's word, "indifferent to") scientists' classificatory practices, some human kinds interact with the classifications scientists are using. Hacking terms these kinds human *interactive kinds* and makes two controversial claims about them: (i) only human kinds are interactive kinds; (ii) human interactive kinds

[1] A slightly modified version of this chapter has been published in the British Journal for the Philosophy of Science under the same title (Laimann forthcoming).

cannot be natural kinds. Both claims have been vehemently criticised – the first on the grounds that there seem to be non-human interactive kinds; the second on the grounds that, even if the phenomenon of interactivity could be limited to human kinds, this would not prevent them from being natural kinds. Despite finding Hacking's detailed case studies insightful, critics have converged on the conclusion that the general account of human interactive kinds which he extracts from them should be rejected.

This chapter aims to challenge this consensus. I argue that, although the critics correctly identify weaknesses in Hacking's argument, the dialectic of the extant debate misses the core conceptual problem of human interactive kinds. The problem is not that these kinds are particularly unstable but "capricious" – their members behave in wayward, unexpected manners which defeats existing theoretical understanding. The reason for that, I argue, is that human interactive kinds are often hybrid kinds consisting of a base kind and an associated status, which makes mechanisms that support patterns of change and stability systematically difficult to understand and predict. Accordingly, a shift in focus is due. I argue that we should stop understanding the question whether human interactive kinds can be natural kinds as hinging on the issue of ontological stability. Instead, we should focus on the role of understanding mechanisms that support patterns of change and stability in our epistemic practices surrounding natural kinds. *Pace* Hacking's critics, considering human interactive kinds from this perspective potentially undermines their status as natural kinds. This has not been acknowledged in the extant discussion and merits further investigation.

In the following section, I start by recapitulating the extant discussion between Hacking and his critics. In Section 4.2, I point out how the dialectic of this discussion centres on the issue of ontological stability over time. I discuss two reasons why this way of framing the debate is misguided. Firstly, it cannot account for the epistemic problems posed by human kinds that participate in stabilizing, as opposed to destabilizing, feedback effects. Secondly, it is based on an oversimplified account of the scientific investigation and use of natural kinds. If these observations are correct, the assumption that human interactive kinds are problematic because their objects are unstable is wrong and has led the discussion astray. In Section 4.3, I argue that human interactive kinds are best understood as hybrid kinds. I then show that such kinds may pose specific difficulties for scientific understanding, which gives credit to the idea that we should exercise special caution in thinking of them as natural kinds.

4.1 THE EXTANT DISCUSSION

4.1.1 Hacking's account of interactive kinds

Hacking's account of interactive kinds is motivated by a number of detailed case studies of psychiatric kinds like multiple personality disorder, child abuse, and schizophrenia (see Hacking 1986; 1988; 1991; 1992; 1995a). Hacking notes that the studied phenomena develop over time in a very peculiar way that is unknown to the natural sciences. The objects of classification "interact" with the classificatory schemes that are used to investigate them: classified individuals change, sometimes up to the point where the original classification is considered obsolete and thus revised. He calls these kinds "interactive" (or "looping") kinds. Phenomena studied in the natural sciences, by contrast, are unresponsive to our classificatory practices. Quarks, to use Hacking's familiar example, do not change in response to how we classify them.

We can understand the underlying process as a two-phase feedback loop. In the first phase, individuals react to the classifications that are (potentially) applied to them by changing their behaviour and characteristics. This phenomenon has been described in the sociological literature on criminal behaviour under the name "labelling theory" (see, for instance, Schur 1971). However, Hacking's account of interactive kinds features a second phase which has not been discussed in labelling theory. He suggests that the changes brought about by labelling can be so extensive as to render the original classification obsolete. Due to labelling effects, individuals might no longer correspond to the criteria or theoretical associations of the original classification. Upon noticing this development, those in charge of the classification (for instance scientists or politicians) may decide that the mismatch is serious enough to necessitate a revision of the definition or theoretical understanding of the classification. Hence, in the second phase, the change in individuals' behaviour or characteristics feeds back into the understanding of the classification used to describe them.

Hacking's discussion of the changing symptom profile of schizophrenia provides a good illustration of this process (Hacking 1999, 113-114). He describes two iterations of the feedback loop, each of which features the two phases described above. According to Hacking, when the diagnosis of schizophrenia was first introduced, experts emphasised "flat affect" and considered auditory hallucinations a minor problem that was not specific to schizophrenia. With auditory hallucinations being such an "unproblematic" symptom, large numbers of people classified as schizophrenic expressed and reported them to their doctors. As a result, auditory hallucinations were found to be universal among

schizophrenics when the classification was operationalised about thirty years later, and were therefore established as a major diagnostic criterion. This is the first iteration of the feedback effect. A second iteration occurred as schizophrenia became a decreasingly "fashionable" diagnosis that individuals tried to avoid. Individuals stopped reporting auditory hallucinations; auditory hallucination ceased to be a widespread characteristic of people diagnosed with schizophrenia, and was successively de-emphasized as a diagnostic criterion.

Hacking makes two controversial claims about interactive kinds. He argues (i) that only human kinds are interactive kinds and (ii) that human interactive kinds are not natural kinds. Some clarifications are in order before we proceed to the criticism of Hacking's account. First, although Hacking often seems to refer to human kinds in general, he is not committed to saying that all human kinds are interactive. To avoid confusion, I will refer to those human kinds which are subject to the feedback effects described above as *human interactive kinds*. Second, given the controversy about the concept of natural kinds, we need to know what concept is at issue in this discussion. Hacking's ideas about natural kinds are sketchy and – including kinds like *mud* (see Hacking 1995b, 352) – unusually permissive.[2] Hacking's critics recognise this, but argue that there is a substantial question as to whether human interactive kinds can be natural kinds according to more orthodox understandings of natural kinds that include biological species as paradigmatic examples (see, for instance, Boyd 1991; Dupré 1993; Millikan 1999). I put aside for now the larger debates about what natural kinds are and whether species qualify, and simply accept the critics' assumption that species are paradigmatic natural kinds. I will come back to the account of natural kinds underlying this debate in Section 4.2.

4.1.2 Classificatory feedback in non-human kinds

Hacking's claims have been subject to extensive criticism. Critics have invoked a variety of non-human kinds which allegedly participate in the same feedback effects as human interactive kind, including kinds of bacteria, marijuana plants, and livestock (Douglas 1986; Bogen 1988; Cooper 2004). The most detailed case has been made with respect to domestic dogs (Khalidi 2010, 345-346). According to Muhammad Khalidi, research suggests that the process by which the species domestic dog diverged from wolves consists of many iterations of the two-phased feedback effect described above. In the first phase, individuals classified as tame were selectively bred, producing increasingly tame

[2] In later work, Hacking (2007) distances himself from the notion of natural kinds altogether, arguing that the concept has outlived its usefulness.

individuals over time. In the second phase, upon recognising that extant individuals do not conform to the existing classification of them, humans revised their classifications (for instance from wolf to domestic dog, and later from domestic dog to particular dog breeds).

These examples are not only used to reject Hacking's first claim that only human kinds can be interactive, but are frequently taken to challenge his second claim that human interactive kinds cannot be natural kinds. As Rachel Cooper points out, many of these examples qualify as natural kinds not only on Hacking's own, somewhat idiosyncratic account, but on many non-essentialist accounts of natural kinds that accommodate species as paradigmatic examples (Cooper 2004, 74-77). Accordingly, it looks like the classificatory feedback effects that Hacking identifies as unique to human kinds in fact produce similar patterns of ontological instability in paradigmatic examples of natural kinds This would imply that both of Hacking's claims are false.

Hacking's staple response to this objection is to insist that the examples above do not qualify as interactive kinds on his view because the objects in question lack awareness of their classification (see, for instance, Hacking 1997, 15). Critics have pointed out a number of problems with this response. First of all, if awareness of one's classification is a necessary feature of interactive kinds, some of Hacking's own examples no longer qualify. Hacking suggests that although young children and individuals with severe autism might be unaware of how they are classified, they might nevertheless participate in classificatory feedback that involves "a larger human unit, for example the family" (Hacking 1995b, 374). The idea seems to be that individuals who are unaware of how they are classified might nevertheless respond to the classification indirectly, for instance by responding to family members or caretakers who are aware of how the individual is classified. This implies that awareness of one's classification is not a necessary feature of interactive kinds.

Second, it has been argued that even if Hacking would consistently restrict his account of interactive kinds to kinds whose members are aware of their classification, he has trouble explaining why these kinds cannot be natural kinds. While change in reaction to becoming aware of one's classification might be specific to humans, it is not clear how this makes human interactive kinds different from the examples of natural kinds discussed above. According to Cooper, in order to make this claim, Hacking would have to assume that classificatory feedback via awareness is of "greater metaphysical significance" than the classificatory feedback we find in other kinds (Cooper 2004, 79). Khalidi makes the same point with respect to feedback effects that are generated phylogenetically, via

selective breeding. He argues that Hacking provides no reason why these phylogenetic feedback effects do not have the same philosophical implications as feedback effects that are created ontogenetically, via awareness (Khalidi 2010, 352).

In other words, both critics agree that even if Hacking stipulatively restricted the concept of interactive kinds to kinds whose members are aware of their classifications, he would still have to face two challenges. First, he would have to exclude some of the examples he previously described as interactive kinds from that category. Second, and more importantly, he would still owe a justification for the claim that human interactive kinds cannot be natural kinds. If Hacking wants to use the notion of interactivity to defend the idea of a fundamental difference between the human sciences and the natural sciences, an *ad hoc* emphasis on awareness will not do. Instead, so the critics suggest, he has to point to an ontological peculiarity of human interactive kinds that disqualifies them as natural kinds. Otherwise, his argument that human interactive kinds cannot be natural kinds fails. I will suggest that these objections, although correct, are somewhat beside the point: their focus on an ontological facet of Hacking's account (instability over time) obscures the main conceptual problems of human interactive kinds. To show this, we need to discuss the premises of the above criticism in more detail, beginning with the underlying account of natural kinds.

4.2 NATURAL KINDS AND ONTOLOGICAL INSTABILITY

What, if anything, could prevent human interactive kinds from being natural kinds? The critics' comparison of human interactive kinds with biological kinds suggest that the difference – if there is one – has to be ontological. This assumption is reflected in Khalidi's question whether human interactive kinds are "real", as well as in Cooper's concern with whether classificatory feedback really marks "a fundamental metaphysical distinction" between human interactive kinds and natural kinds. However, when we look at how both critics frame their investigation, a different aspect emerges. Cooper motivates her discussion with reference to the central epistemic role that natural kinds play in scientific inquiry:

> If human kinds are natural kinds then this suggests that accounts of laws, explanations, and the basis of sound inductive inferences, developed for the natural sciences, can be carried across into the human sciences. If human kinds are not natural kinds, then this will be a reason for thinking that distinct accounts will be required.
>
> (Cooper 2004, 84)

Similarly, Khalidi suggests that we should consider human interactive kinds as real, via adopting "a weak realist view that considers as real any kind that plays an indispensable role in explaining phenomena, making successful predictions, and otherwise featuring in successful inductive inference" (Khalidi 2010, 358). Both remarks suggest that the guiding motivation of the debate is not purely metaphysical interest, but the question whether human interactive kinds can fulfil the epistemic role of natural kinds.[3] The critics' concern with the status of human interactive kinds as natural kinds is effectively an epistemological and methodological one: if human interactive kinds are natural kinds, we do not need to come up with radically new approaches to understand them – their investigation can simply be modelled on the methods and epistemic practices used in the natural sciences.

This hope stands in sharp contrast with some of Hacking's remarks. He suggests that any attempt at investigating human interactive kinds in the same way as natural kinds is destined to fail, and that more suitable approaches are yet to be invented (see, for instance, Hacking 1997). Against this background, we can understand the rejection of Hacking's account as an attempt to reassure us that the phenomenon Hacking describes is not as epistemically troublesome as he makes it out to be. To evaluate Hacking's claims, we need to understand what could possibly hinder human interactive kinds from being scientifically investigated and epistemically used in the same way as natural kinds.

On many occasions, Hacking suggests that the problem with using human interactive kinds as natural kinds has to do with the fact that they are unstable. In Hacking's words, human interactive kinds are "on the move" or "moving targets" (see, for instance, Hacking 1999, Chpt. 4; 2006). This idea resonates with the example of schizophrenia discussed in Section 4.1. There, it seemed that by classifying individuals as schizophrenic, investigators unleashed a process in which the classified individuals change until they no longer fit the original classification. The resulting epistemic problem seems to be described most clearly with respect to the kind *child abuse*. Here, Hacking suggests that there might not be "a stable object […] to have knowledge about" (Hacking 1995a, 61). The idea seems to be that members of human interactive kinds constantly change in virtue of feedback effects, and we are not able to acquire knowledge and make inductive inferences about objects which constantly change over time. Accordingly, Hacking's critics have focussed on instability as a potential problem for human interactive kinds' status as natural kinds. Khalidi, for example, suggests that human interactive kinds seem

[3] Cooper and Khalidi develop these accounts in more detail elsewhere (see Cooper 2005, Chpt. 2; 2007, Chpt. 4; Khalidi 2013).

to pose an epistemological problem because "after successive iterations of the looping effect, it seems that we may no longer be dealing with the same thing we started with" (Khalidi 2010, 342).

In other words, the debate is essentially about whether the members of human interactive kinds are unstable in a way that precludes them from functioning epistemically as natural kinds. Hacking seems to affirm this claim. His critics reject the claim on the grounds that similar patterns of instability are not considered a problem in the many examples of non-human kinds presented above. Neither side of this debate, however, seems to consider the association between ontological stability and the epistemic role of natural kinds worthy of further scrutiny. In the following, I discuss two reasons for questioning this assumption. Firstly, it is based on an account of natural kinds as vectors for projections and generalisations that is oversimplified. Secondly, it cannot account for the epistemic problems posed by human kinds that participate in stabilizing, as opposed to destabilizing, feedback effects.

4.2.1 Understanding instability

In order to bring into focus the assumptions about the relation between ontological stability and the epistemic features of natural kinds that form the background of the above discussion, we need to specify what kind of instability is considered a potential threat to natural kind status, and why. For that purpose, we first need to specify what sort of change we are talking about. As described above, there are two sorts of change involved in the classificatory feedback that characterises human interactive kinds. There can be changes to the members of a kind, for instance when the extension of the kind changes (new members join, extant members lose membership or cease to exist), or when the characteristics of the individuals within that extension change (members acquire new properties or shed old ones). Alternatively, there can be a change in the theoretical beliefs associated with the kind, such as when we discover new properties of the members and adapt our theoretical understanding to accommodate these. Although participants in the debate occasionally talk of kinds themselves "changing" or "being unstable", this terminology should be avoided because it is ambiguous between these two quite different processes: the change of members is something that happens in the world, the change of theoretical understanding is something we deliberately bring about. What participants in the debate mean when they talk of a kind being "unstable" is that the members of the kind change in ways that require us to alter our existing theoretical understanding of the kind.

Note that not just any type of change among members constitutes this sort of instability. Change is abundant in the natural world and scientists understand, explain, and predict the behaviour of a great variety of objects which change over time, such as reactive chemical compounds, or animals that undergo metamorphosis. Take the kind *water* (H_2O). We know a lot about the properties of this kind, for example that it has a melting point of 0 °C and a boiling point of 100 °C. However, we do not think that these properties are fixed or absolute, but know that they change depending on atmospheric pressure. Accordingly, natural kinds can have properties which are theorized as changing under specific circumstances, just as the melting point and boiling point of water are theorized as changing relative to atmospheric pressure. Therefore, what we mean when we say the natural kind water is stable is not that instances of water do not change under differing circumstances. We mean that, over time, instances of water do not change or develop new properties that are at odds with our existing scientific understanding of water. This suggests that we need to be more precise when asking whether instability prevents a kind from functioning as a natural kind category. The problem with human interactive kinds is not merely that the classified objects change, but that they change in ways which are unforeseen by our extant theoretical understanding of the kind. This is not the case for chemical compounds like H_2O.

The case is different for biological kinds like species. Here, existing members of a kind are constantly replaced by new members with slightly different properties. As a result, the set of properties that characterises members of a species can be transformed over time—instances of domestic dog today are characterised by very different properties than instances of domestic dog 200 years ago. Hence, instances of a species can, in a sense, change properties in a way that is at odds with our existing understanding of the species at any given point. When critics liken the instability of a human interactive kind like schizophrenia to the instability of biological kinds like domestic dog, what they have in mind is this instability over time of the set of properties associated with a kind. The rich biological literature on species like domestic dog suggests that biological kinds are quite capable of facilitating prediction, explanation, and inductive inference, and thus epistemically qualify as natural kinds.

Since members of human interactive kinds seem to change over time in much the same way as biological kinds, Hacking's critics conclude that it is implausible to claim that the latter can have natural kind status whereas the former cannot. They anticipate that Hacking might respond by arguing that members of human interactive kinds change at a

significantly higher rate than members of biological kinds, and cannot have natural kind status for that reason. However, Cooper and Khalidi dismiss this point fairly quickly (Cooper 2004, 79; Khalidi 2010, 350). They argue that even if it was evidently true that the members of human interactive kinds change faster than those of non-human kinds – which they doubt – this would not by itself explain why human interactive kinds cannot be natural kind categories. The difference is, after all, only one of degree.

But at this point, it seems like the critics' metaphysical concerns with natural kinds have gotten ahead of their underlying epistemic motivations. It might be plausible to argue that a gradual difference in the rate of change cannot establish a *metaphysical* difference between human interactive kinds and natural kinds. However, given the motivating *epistemic* concern with natural kinds, the dismissal seems somewhat hasty. From an epistemic perspective, the claim that human interactive kinds cannot function as natural kinds because they change too quickly deserves serious consideration. After all, it seems perfectly reasonable to assume that a classification's ability to facilitate inductive inferences that allow us explain the behaviour of past instances and predict the behaviour of future ones depends crucially on how much its objects have changed in the meantime.

A defender of Hacking could develop this point further by arguing that an epistemically significant threshold lies between the rates of change of members of biological kinds and those of human interactive kinds: while members of biological kinds change slowly enough for our scientific understanding to catch up, members of human interactive kinds outrun our efforts to theorise about them. Ron Mallon explores this idea in some detail (Mallon 2016, Chpt. 7).[4] According to Mallon, whether we can have knowledge about a human interactive kind depends on whether scientists manage to increase the accuracy of their theories about members of the kind at a higher rate than the rate at which the members change. I call this the *hare-and-tortoise* account of scientific understanding. Mallon illustrates this account in the case of biological species, arguing that scientists

> can have knowledge of members of these changing kinds that allows us to engage in successful induction, prediction, explanation, and intervention because our capacity to gain accurate knowledge of these kinds can (sometimes) be far more rapid than the processes that underwrite biological change.
>
> (Mallon 2016, 166)

[4] Interestingly, Mallon uses this proposal to defend rather than challenge the claim that human interactive kinds can function as natural kinds. He suggests that we should expect human interactive kinds to often develop at a slower rate than the theories we formulate to explain them, because stabilizing feedback tends to be more prevalent and powerful than destabilizing feedback (see Mallon 2016, 173-181).

Certain aspects would need to be addressed further to develop this idea into a solid argument – for instance how to operationalize rates of change and rates of theory improvement in a way that allows us to compare the two. But instead of doing that, I want to draw attention to the limitations of the accounts of natural kinds and scientific understanding that underpin this line of argument.

To begin with, the hare-and-tortoise account might suggest that there is an inverse relationship between the objects' rate of change on the one side, and our ability to develop scientific understanding of them as natural kinds on the other: the more idle the objects of inquiry, the better they can be studied and function as natural kinds. However, there are reasons to think that change at a very slow pace poses problems of its own. Picking up Mallon's example of species, it would not be far-fetched to suggest that the slow rate at which most readily observable species evolve has hindered our understanding of evolution. If horses and birds had the generation time of bacteria, we might have arrived at a theory of evolution, and hence a better understanding of the natural kinds horse and bird, at a much earlier point in human history. Change at a very slow rate tends to escape our attention and if this happens, we fail to incorporate this aspect into our theoretical understanding of the kind. Admittedly, the relative stability of the members of many species has epistemic advantages: we can make a great number of predictions and inductive inferences about members of the kind, precisely because change occurs at a rate slow enough as to not interfere with them. However, our inductive inferences across wider time spans will be susceptible to error, and our explanations will lack information on phylogenetic history and evolutionary mechanisms. Overall, we would be inclined to say that, without these, our knowledge of the kinds in question is highly incomplete at best.

The example above shows that a slow rate of change of the members of a kind is by no means sufficient for the kind to facilitate scientific understanding. Other examples suggest that a relatively slow rate of change is not necessary for acquiring scientific understanding either. Consider bacteria. For some strains of bacteria, an individual can within thirty hours grow into a population in which every single base pair in the genome has mutated thirty times.[5] It seems unlikely that scientific theories about bacteria really approach accuracy at a faster rate than that. Fortunately, scientists working on these organisms do not start out from scratch, but can draw on theoretical resources from other

[5] See Pray (2008).

areas. For example, much of the knowledge applicable to bacteria is derived from the study of species which change at a less breath-taking speed, such as fruit flies. Additionally, experimental setups can be used to limit possible causes of change and to ease the process of tracking members of a specific strain without having to identify each bacterium on the basis of shared characteristics, as was achieved by the development of the pure culture method in microbiology (see O'Malley 2014).

These arguments suggest that the hare-and-tortoise account that motivates the focus on instability is overly simplistic. Scientists' ability to improve the accuracy of their theories does not simply stand in inverse relationship to the studied objects' rate of change, but depends on a host of factors, such as the possibility of making relevant observations, the ability to draw on an existing understanding of underlying mechanisms, and the opportunity to study objects under laboratory conditions. Accordingly, when deciding how well human interactive kinds can fulfil the epistemic role of natural kind categories, all these factors need to be taken into consideration. This point has not been explicitly addressed in the extant discussion on human interactive kinds, which focusses mainly on stability.

4.2.2 The problem of stabilizing feedback

The second problem with focussing on ontological stability as a crucial feature of natural kinds is that this view cannot account for the epistemic challenges posed by human classifications which are stabilized, rather than destabilized, by classificatory feedback. Hacking tends to focus on case studies where classificatory feedback makes individuals "outgrow" existing classifications, such as the example of schizophrenia discussed above. Call this type of classificatory feedback *destabilizing* feedback. However, there is a second type of classificatory feedback – *stabilizing* feedback – which achieves the contrary result: labelling effects reinforce properties associated with a classification, which is then interpreted as support for the existing classificatory practice. Standard examples in labelling theory describe such a process. They suggest, for instance, that the fact that someone has been labelled a criminal plays a role in their engaging in further criminal behaviour (see, for instance, Lemert 1951; Becker 1963; Chiricos et al. 2007; Worrall & Morris 2011). If the confirming labelling effects of a particular category are powerful enough, members of the category will generally conform to the properties associated with the category to a higher degree than they would have had, had they not been labelled. In response, those in charge of the classification might interpret the fact that individuals fit their labels so neatly as confirmation of the classificatory practice. In keeping with

Hacking's metaphor, we might say that human kinds which are subject to stabilizing feedback are "held in place" rather than "sent on the move".

For someone who believes that ontological instability is the main threat to human interactive kinds' status as natural kinds, stabilizing and destabilizing feedback effects need to be treated radically differently. While destabilizing feedback prevents human kinds from being natural kinds, stabilizing feedback would presumably make them more suitable candidates for natural kind status. After all, if natural kind categories need to refer to stable objects in order to facilitate induction, explanation, and prediction, and stabilizing feedback provides us with such stable objects, it should enable at least some human interactive kinds to function as natural kinds.

Dominic Murphy makes an argument along these lines (Murphy 2006, 267-270). He suggests that if the norms, social pressures, stereotypes, or medical opinions that facilitate stabilizing feedback persist over time, the resulting patterns of behaviour that characterise a human interactive kind might "freeze in place", thus making the kind perfectly suitable for inductive inferences. Accordingly, a proponent of the view that ontological instability is the main threat to natural kind status would have to hold one of the following claims: (i) the concept of human interactive kinds includes only kinds which are subject to destabilizing feedback, or (ii) the concept of human interactive kinds also includes kinds which are subject to stabilizing feedback, but this does not commit us to saying that the latter cannot be natural kinds. While Hacking's position on the matter is not entirely clear, from an epistemological perspective, both of the above claims should be rejected.[6] The reason for this is that the epistemic challenges posed by stabilizing feedback can be substantial, and are in some respects more detrimental to the acquisition of scientific knowledge than the challenges associated with destabilizing feedback.

The debate on the causes of differences between men and women is a notorious case in point. As already noted in John Stuart Mill's *The Subjection of Women*, the crux in this debate is that, for many observed behavioural or psychological differences between men and women, we have trouble identifying whether they are due to "nature" or due to "society" (Mill 1869/1984). In other words, it is difficult to identify whether the observed differences are due to underlying natural, biological differences between men and women or due to differences in social upbringing and differential social constraints and opportunities. If the latter factors play a role (as we now have plenty of evidence to

[6] Hence, I am not suggesting here that Hacking's critics are guilty of misinterpretation by wrongly attributing to him either (i) or (ii). At least with respect to (i), careful readers will find passages that support it, as well as passages that undermine it (see Hacking 1999, 34 versus Hacking 1995b, 369-370).

believe), it is very compelling to think of men and women as human interactive kinds that are subject to stabilizing feedback effects.

We can imagine the underlying two-part feedback mechanisms operating in the following way: In the first part, individuals are born into a society that has certain preconceived ideas about men and women (for instance that there are natural differences between them which not only determine their distinct morphological features, but also differences in character, abilities, and preferences). The society socialises individuals and arranges social institutions in accordance with these preconceived ideas. As a result, individuals classified as men or women continuously encounter differential social expectations and constraints and, over time, develop behaviour patterns, character traits, and abilities suitable to their circumstances – they come to fit their classification. In the second part, the fact that individuals classified as men or women squarely conform to these preconceived understandings is interpreted as evidence for the adequacy of the existing classificatory practice and its theoretical associations. It looks like men and women do naturally differ in character, ability, and preferences. This feedback mechanism is iterated as scientific testimony to the existence of such natural differences between men and women emerges. Scientific testimony strengthens the associated labelling effects, which is again, in turn, interpreted as confirmation of the classificatory practice and the theoretical understanding that underpins it. Due to these classificatory feedback effects, scientists came to firmly understand men and women as natural kinds that facilitate explanation and prediction not only of anatomical features, but also of a broad range of behavioural and psychological characteristics.

Assuming that this story is more or less accurate, we can see how stabilizing feedback effects not only obscured and facilitated the oppression of women, but also contributed to an erroneous understanding of the kinds *men* and *women*.[7] Many explanations facilitated by this understanding have been either false or substantially incomplete. Moreover, since the theoretical understanding suggested that differences between men and women are largely invariable across different societies, predictions and inductive inferences made on its basis were unreliable. In other words, the example above suggests that human kinds which are subject to stabilizing feedback can make for very poor natural kinds.

But more than that, there is reason to believe that human kinds which are subject to stabilizing feedback are, in some respects, worse candidates for natural kind status than

[7] This is not to say that stabilizing feedback alone is responsible for the poor epistemic outcome. Other factors, such as bias on the part of an overwhelmingly male research community, have arguably played an important role (see, for instance, Longino 1990).

human kinds which are subject to destabilizing feedback. Destabilizing feedback is, in some sense, transparent. The fact that classified phenomena resist and undermine our classificatory practices rubs our nose in the fact that the classifications we are using are based on an inadequate understanding of the phenomena in question. Stabilizing feedback, by contrast, is opaque. The apparent success of our classification can lull us into a false sense of security about the adequacy of the theoretical understanding that underpins the classificatory practice. If these observations are correct, and stabilizing feedback is at least as, and arguably more, epistemically challenging than destabilizing feedback, the assumption that human interactive kinds are problematic because their objects are unstable is wrong. Instead, the case of gender differences suggests that the problem is down to an inadequate understanding of the underlying determinants of change and stability in members of the kinds—only when we understand the mechanisms that support patterns of change and stability among the members of a kind are we in a position to provide accurate explanations and make inductive inferences across a variety of contexts.

4.2.3 Summary

Putting together the observations from the previous sections, the assumed connection between ontological stability and the epistemic features of natural kinds starts to look rather fragile. There is reason to believe that the focus on ontological stability reflects an overly simplistic hare-and-tortoise account of scientific inquiry and natural kinds. The case of stabilizing feedback corroborates these findings. It suggests that using ontological stability as a chief criterion for natural kind status may leave us with an epistemically thin and potentially misleading understanding of the kinds in question. Fortunately, it also indicates where a more nuanced understanding can be found: our epistemic practices surrounding natural kinds require knowledge of the causal processes that support patterns of change and stability in the classified objects. In order to be able to explain, predict, and make inductive inferences about the behaviour of members of a kind, we not only need to know *that* members typically display certain patterns, but also *why* they display these patterns or *what* produces them. In other words, natural kind categories should be understood not simply as vectors for projections and generalisations, but as analytic tools that incorporate assumptions about the causal mechanisms which constitute the kind.

These insights apply neatly to the example of domestic dog we started out with. Proponents of the hare-and-tortoise account suggest that domestic dog qualifies as a natural kinds because change in the set of properties associated with this kind occurs at a

pace slow enough for our understanding to "catch up" and produce accurate explanation and predictions. The discussion above suggests that something different is going on. It suggests that domestic dog is a natural kind because we understand sufficiently well the evolutionary mechanisms by which members of the kind change their characteristics over phylogenetic time. Hence, while changes in the set of properties associated with this kind might, in one sense, overhaul our existing understanding of domestic dog – dogs in 200 years will probably look very different from dogs today – it is, in a different sense, perfectly in accord with our existing understanding. By contrast, if Hacking's description of the historical development of schizophrenia is correct, the reason we are taken aback by the instability of the set of properties associated with schizophrenia is that we have a wrong or incomplete understanding of the causal processes that support it.

In other words, in trying to understand whether human interactive kinds can be natural kinds, we ought to stop putting so much emphasis on stability and instead ask if there is anything about these kinds that hampers our efforts to understand the underlying causal processes. In the following section, I argue that considering human interactive kinds from this perspective provides some reasons to be cautious about their status as natural kinds, thus rendering Hacking's account more convincing than his critics acknowledge.

4.3 CAPRICIOUS KINDS

What, then, is the problem with human interactive kinds, if not unusual instability? I suggest that the problem has to do with their peculiar ontological structure. Human interactive kinds tend to have a dual nature: while we commonly think of human interactive kinds in terms of the properties that explicitly define the category, they can also be understood in terms of the social position that individuals occupy in virtue of being recognised as members of the category. In other words, human interactive kinds fit the hybrid kind model developed in Chapter 1. They consist of a base kind, constituted by the properties that define the category, and an associated status kind, constituted by the social position that individuals acquire *qua* being recognised and treated as members of the specific category.

The example of men and women from the previous section is useful to illustrate this idea. It is one of the few cases where the dual nature of a hybrid kind has been comprehensively conceptualised, in the form of the sex/gender distinction. Feminists have historically used the sex/gender distinction to tackle the idea that differences between men and women are biologically determined (see Mikkola 2017). Roughly

speaking, the distinction between sex and gender was meant to distinguish differences in biology (*sex*) from differences that are due to culture and society (*gender*). Terminologically, this distinction is sometimes expressed by using "male"/"female" to refer to sex categories, and "men"/"women" to refer to gender categories. I do not adhere to this terminology, but instead use "men" and "women" in the theoretically naïve sense that makes no such explicit distinction.[8]

While there are many ways to spell out the idea of gender (for instance in terms of gender identity, or socialised behaviour), the understanding which is relevant to my idea of a hybrid kind is best captured by the feminist slogan "gender is the social meaning of sex". This slogan expresses the idea that gender is a social position or role that individuals occupy in virtue of being recognised as members of a specific sex, an idea which has been developed in much detail by Sally Haslanger (2012) and Asta Sveinsdottir (2011; 2013). As a social position, gender is characterised by the norms, expectations, privileges, constraints, and opportunities that apply to individuals *qua* being recognised as members of a certain sex. In my terminology, sex is the base kinds, and gender (understood as a social position) the associated status kind.

As Asta (2013) argues in detail, the relationship between membership in the base kind and membership in the status kind is of a special and somewhat fragile nature – members of the base kind come to occupy the social position that characterises the status kind only if they are recognised as members of the base kind, and individuals who are wrongly believed to be members of the base kind might nevertheless come to occupy the associated social position. Although this relationship does not guarantee complete coextension of the base kind and the status kind, the properties of the base kind and the properties of the status kind are associated reliably enough to suggest that the terms "men" and "women" refer to hybrid kinds – they are commonly understood as, and often succeed in, distinguishing people on the basis of biological characteristics, yet they also unwittingly track an associated distinction in terms of social position.[9] On this account, the distinction between sex and gender can be understood as an attempt to conceptualise the hybrid nature of the human categories men and women, with *sex* denoting the base kind and *gender* the associated status kind.

While Haslanger and Asta use this perspective primarily to develop a detailed metaphysical understanding of gender and other status kinds, I am more interested in

[8] See Saul (2006) for an argument that ordinary speakers do not distinguish sex from gender.

[9] Note that Haslanger and Asta would probably disagree with this characterisation—they suggest that "men" and "women" should better be understood as referring exclusively to the associated status kinds. See Saul (2006) for a discussion.

what it tells us about the prospect of using human interactive kinds as natural kinds. I think the classificatory feedback effects described by Hacking can be understood as feedback effects between a base kind and the respective status kind. By being classified as members of a human category defined in terms of certain base properties, individuals come to occupy a specific social position (become members of the corresponding status kind) that is characterised by specific norms, expectations, constraints and opportunities, and that influences how others relate to them as well as how classified individuals relate to themselves. In virtue of these features, membership in the status kind can affect the characteristics of classified individuals, which may stabilize or destabilize our theoretical understanding of the base kind. In the remainder of the chapter, I argue that understanding human interactive kinds as hybrid kinds should make us wary about treating them as natural kinds. The reason for this is that hybrid kinds are susceptible to two problems which complicate their functioning as natural kinds: (i) biased conceptualisation, which theorises about the base kind whilst disregarding the status that is imposed onto members of the base kind; and (ii) difficulty conceptualising, explaining and predicting the social status that is associated with a base kind.

4.3.1 Biased conceptualisation

Biased conceptualisation describes a phenomenon by which we theorise about and investigate the base of a hybrid kind while paying little attention to the associated status. The discussion of stabilizing feedback suggests men and women had been conceptualised in a biased manner before the distinction between sex and gender was introduced. Similarly, reconsider Hacking's paradigmatic example of schizophrenia. Schizophrenia is commonly understood either in terms of a specific symptom profile, or in terms of an underlying neurological condition that is assumed to produce these specific symptoms (Murphy 2006). Yet the category schizophrenia also picks out a status kind, which is a specific position in a network of social relations that individuals occupy in virtue of being classified as schizophrenic. Hacking's discussion details how people diagnosed as schizophrenic are singled out for particular interactions and treatments, and are subject to a number of specific expectations, opportunities, and constraints. In fact, Hacking makes quite clear that it is this network of social relations which mediates classificatory feedback in the kind *schizophrenia*.

How does biased conceptualisation threaten the natural kind status of human interactive kinds? In order for a kind to function as a natural kind, we need to have a sound theoretical understanding of the causal processes that underpin the properties

associated with it. We want to know not only that members of the category typically behave in certain ways, but also why they typically behave in these ways and under which circumstances we should expect them to behave differently. However, when we conceive of a human interactive kind solely in terms of the base kind, without considering the associated status, causal pathways associated with the status disappear out of sight. If these causal pathways have a significant influence on the properties of classified individuals, undetected biased conceptualisation will prevent us from developing the causal understanding that is necessary to use human interactive kinds as natural kinds. In other words, although biased conceptualisation does not necessarily affect all human interactive kinds, or necessarily preclude a proper understanding of all those kinds affected by it, it is an unacknowledged potential hindrance to using human interactive kinds as natural kinds and, as such, needs to be addressed in the debate.

The previous discussion suggests that the categories men and women have been severely affected by biased conceptualisation. Here, the focus on a biological conceptualisation concealed the role social positioning played in producing observed differences. Accordingly, scientists prioritised the search for biological determinants of the observed differences (such as brain size and shape, or hormones) over the search for social ones (such as socialisation or social structural constraints). The same might have been true for schizophrenia, if Hacking's description is correct and changing medical beliefs did play a significant role in the changing symptom profile. In this case, it seems that conceptualising schizophrenia as a cluster of symptoms or as a neurological disorder, without taking into account the associated status, obscured changes in medical beliefs about schizophrenia as a possible cause for the changing symptom profile.

In several places, Hacking remarks on our tendency to "biologize" or "geneticize" human interactive kinds (see, for instance, Hacking 2006; 1995b, 353). Although these remarks resonate somewhat with my idea of biased conceptualisation, they are misleading in that they suggest that biased conceptualisation always necessarily involves human kinds which are conceptualised as biological. This is not the case—kinds which are explicitly conceptualised as social can be hybrid kinds affected by biased conceptualisation, too. Other passages in Hacking align with this idea. He cautions that the classification woman refugee is associated with social and material factors that affect the characteristics of women thus classified (1999, 10-11), and that our tendency to think of children who watch television as a "species", might reify the kind child viewer of television via classificatory feedback effects (1999, 27).

Unfortunately, Hacking does not say anything more concrete about effects of biased conceptualisation in each case. But we can illustrate the idea with the example of the kind *unemployed*. Here, the base kind, understood as being without paid work but available to work, is explicitly defined with respect to social institutions. Nevertheless, being unemployed is also associated with a status which, among other things, involves social stigma. If the social stigma of people classified as unemployed has a crucial influence on their properties, biased conceptualisation that only considers the base kind could lead to gaps in our understanding.

There is some evidence that this has taken place with respect to health disparities between employed and unemployed people. O'Donnell et al. (2015), for instance, found evidence that stigma negatively affects the psychological and physical health of unemployed people, but also note that there is very little existing research on this hypothesis. Former studies, they argue, instead focus on factors like financial strain, or lack of time structure, social contact, and activity—all factors typically associated with the base kind rather than the status of the hybrid kind unemployed. As O'Donnell et al. observe, this perspective not only provides a limited theoretical understanding of the existing health disparities, it also obscures potential interventions, such as changing public perceptions of unemployment or teaching skills for coping with stigmatization.

4.3.2 Studying social status

If biased conceptualisation was the only potential problem with using human interactive kinds as natural kinds, the solution would be fairly straightforward: simply identify the associated status and understand what feedback effects it has on classified individuals. Unfortunately, there are reasons to believe that understanding these statuses and their feedback effects is anything but straightforward. As Jaakko Kuorikoski and Samuli Pöyhönen point out, although much of social science is limited to describing patterns of social life, only an understanding of the underlying mechanisms allows scientists to make inferences about counterfactual scenarios and enables them to extrapolate findings to new contexts and identify effective interventions (Kuorikoski & Pöyhönen 2012, 191). Hence, in order to be able to explain and predict how the status associated with a classification affects classified individuals, we need to understand not only the social and psychological mechanisms that mediate feedback effects, but also the mechanisms that stabilise and modify the status over time. There are several factors that potentially complicate this understanding.

Consider first the feedback-mediating mechanisms. Several philosophers have provided extensive discussions of these mechanism, often illustrated with examples supported by social scientific research (Drabek 2014; Kuorikoski & Pöyhönen 2012; Mallon 2016; Murphy 2006). Accordingly, I will not repeat their points here, but simply point to the diversity of causal pathways that this literature has identified. Mallon, for instance, distinguishes three main pathways by which classifications can lead classified individuals to change their behaviour: intentional change of behaviour, automatic change of behaviour, and environmental construction (Mallon 2016, 68-89). Each of these, he suggests, can occur via several different causal pathways. Intentional change, for instance, can happen via change in salient possibilities for action, or via strategic or non-strategic reasoning. Environmental construction involves processes such as transmission of culture and institutions, or modifications of the material and spatial environment. In addition to that, Kuorikoski and Pöyhönen point out that classificatory feedback can operate with or without the individual being aware of the classification (Kuorikoski & Pöyhönen 2012, 196-197). They discuss examples showing how classificatory feedback can happen without awareness, through processes such as the alteration of the practical reasoning of classified individuals or the modification of other people's expectations towards classified individuals.

This literature suggests that, although it is possible to identify and, to some extent, empirically investigate the social mechanisms that mediate classificatory feedback, these mechanisms are quite varied and complex. To complicate things further, different mechanisms may pull in different directions, thus amplifying or attenuating their respective effects. For example, on the intentional pathway, the effect of me being classified as a criminal might be that the classification troubles me, so that I resolve to make special efforts to act lawfully in the future. At the same time, my efforts to do so might be frustrated by the structural and material constraints that affect me as someone classified as a criminal. I might, for instance, no longer be eligible for a variety of jobs, which makes earning a living via illicit activities a more compelling option.

In addition to that, attempts at understanding and predicting classificatory feedback are complicated even further by the fact that the social meanings associated with classifications may vary both synchronically and diachronically. At any given time, a classification can mean different things in different contexts and interact with other classifications. Consider again the example of men and women. In this case, the conceptualisation and investigation of the associated status is relatively advanced, arguably

more so than in any other human interactive kind. In the last few decades, a comprehensive literature that theorises gender as a status kind has emerged (see, for instance, Oakley 1972; MacKinnon 1989), followed by systematic empirical investigation into the associated determinants of differences between men and women. Fields like social psychology, for instance, now provide ample empirical support for activists' and critical theorists' long-held claim that psychological differences between men and women cannot be explained purely in terms of biology, but require consideration of their differing treatment and positioning in society (see, for instance, Eccles 1987; Eagly 1987; West & Zimmerman 1987; Spencer *et al.* 1999).

However, simultaneously with these developments, a discussion has emerged as to whether a unitary category of women's gender is a useful category at all, given how racial, cultural, and class differences influence the positioning and experiences of individuals classified as women (see Spelman 1988; Crenshaw 1991; Butler 1990; Mikkola 2006; Stoljar 2011). At the centre of this discussion is the observation that the specific social position that an individual occupies in virtue of being classified as a woman varies greatly depending on a number of other factors. These include the background culture in which the classification is used, other classifications that are applied to the individual, as well as not classification-induced social and economic factors.[10] The debate suggests that it might not always be possible to identify a unitary status associated with a certain classification. Instead, in order to understand the causal processes that support a human interactive kind, one needs to understand how the classification affects individuals in different circumstances and in interaction with other classifications.

In addition to that, the status associated with a classification may change over time. Again, the problem is not that the social meanings of classifications change at all, but that they change over time in ways that are difficult to explain and predict. Why did the Stonewall riots in 1969 in New York lead to a gay liberation movement that radically changed the status kind associated with the category homosexual? Historians can discuss the merits of different hypothesis to explain this event and its impact, but they have little way to empirically decide between them. Events like the rise of the gay liberation movement are the result of complex social and political processes that possibly involved a unique constellation of a myriad of factors that cannot be reproduced or tested under

[10] By "not classification-induced factors" I mean factors that do not depend on the individual being recognised as of a certain kind, although the factors might be causally associated with a certain kind. For instance, many people are poor because they are working class, but their being poor is not (or not primarily) due to being classified as working class.

laboratory conditions. As a result, social scientists cannot explain or predict changes of the meanings associated with human kinds with any certainty.

These are both familiar points in the discussion of social scientific methodology, yet their relevance to the question whether human interactive kinds can function as natural kinds has not been explicitly addressed in the extant literature. In particular, they suggest that status kinds themselves may often make poor candidates for natural kinds. If we cannot explain and predict the social meanings associated with human classifications, we are in no good position to explain their respective classificatory feedback effects, or to make reliable inferences about what feedback effects the classification is going to bring about under different circumstances. Hence, although the above discussion does not establish that human interactive kinds can never function as natural kind categories – there might be cases where we have a firm understanding of the associated status, and the mechanisms facilitating feedback are few and well-studied – it does provide some reasons to be cautious.

4.4 CONCLUSION

In this chapter, I discussed Hacking's heavily criticised suggestion that human interactive kinds cannot be natural kinds. I suggested that there might be more to Hacking's claim than his critics acknowledge, albeit not for the reasons Hacking identifies. Hacking suggests that interactivity is primarily a phenomenon of instability of the set of properties associated with a kind. His critics rightly object that interactivity thus understood does not preclude human interactive kinds from being natural kinds. I argued that both sides miss the core threats to natural kind status because they presuppose an oversimplified understanding of the epistemic role of natural kinds. Natural kinds are not simply vectors for projections and generalisations, but analytic tools that incorporate assumptions about the causal mechanisms which constitute the kind. At the same time, human interactive kinds tend to have an ontological structure which compromises their ability to fulfil this epistemic role. They can often be understood as hybrid kinds, consisting of a base kind and an associated status kind, and are subject to several features that potentially threaten their status as natural kinds. These include the tendency towards biased conceptualisation, the diversity and complexity of mechanisms mediating classificatory feedback, and most importantly, the fact that there is reason to think that status kinds themselves make poor candidates for natural kinds.

What are the methodological implications of my account? Recall that the discussion so far has been characterised by two methodological positions. According to Hacking, the phenomenon of human interactive kinds supports the difference thesis. For him, the fact that human interactive kinds cannot be natural kinds implies that we need radically new and different methods for understanding these kinds. His critics, by contrast, seem to support the unity thesis. By insisting that human interactive kinds can be natural kinds, they suggest that investigating these kinds is just "science as usual" – we do not need methods that radically differ from those of the natural sciences.

My own account locates the truth somewhere in between these two positions. Although there might be cases in which we understand the associated status and its feedback effects well enough to use a human interactive kind as a natural kind, there is reason to believe that some human interactive kinds will be unsuitable as natural kinds. Yet this need not imply that investigating these kinds requires a radically new methodology. Coming back to the example of the kinds *men* and *women*, extant work in this area suggests that many researchers are perfectly well aware of the challenges and have found different ways of responding to them. After the crucial initial step of theoretically distinguishing the status kind gender from the base kind sex, feminist theorists have offered an understanding of gender as diverse and context-specific (see, for instance, Spelman 1998; Butler 1990), or suggested to understand gender along a specific politically relevant dimension (see, for instance, MacKinnon 1989; Haslanger 2012). These accounts of gender might not have (and are often not intended to have) the inductive power that we typically associate with natural kinds. But they might nevertheless provide an adequate understanding of how particular social mechanisms produce properties associated with *men* and *women* in specific contexts, or elucidate aspects of gender that are of central importance in emancipatory politics. In other words, contrary to Hacking's claim, the challenges of human interactive kinds need not demand a radically new scientific methodology. In many cases, a better engagement with the resources that are already on offer will do.

5

WHAT CAUSES GENDER DIFFERENCES?

HYBRID KINDS AND CAUSAL EXPLANATION

Our inquiry so far helped us gain a better understanding of the relationship between ontological questions in the social sciences on the one hand and the natural sciences on the other. In particular, it allowed us to see that the hybrid nature of many human classifications has interesting epistemic implications. It is due to their structure of base and associated status that kinds like *women*, *homosexual*, or *schizophrenia* can simultaneously be the subject of natural scientific and social scientific inquiry. Against this background, it is time to consider a different set of questions: How do the explanations that are offered by social scientists and natural scientist in such a constellation relate? Do the explanations provided by each party have to be integrated in order to gain a fuller understanding of the phenomena under consideration? What role do explicitly moral or political aims and considerations play in this context?

In order to answer these questions, I will turn to one of the most contested issues between natural and social scientists: the debate on explanations of gender differences.[1] I will focus specifically on two generally opposed parties in this debate. These parties are proponents of evolutionary psychological explanations on the one hand, and proponents of social scientific explanations – in particular those arguing in the context of an explicit commitment to feminism – on the other. Broadly speaking, both approaches are concerned with psychological and behavioural differences between men and women. At the same time, they tend to conceptualise their subject matter in strikingly different ways, each reflecting the conventions and interests of their own discipline. Evolutionary psychologists mainly understand men and women as different types of biological organisms which have been subject to distinct evolutionary pressures. Social scientists, by contrast, look at men and women primarily as groups of individuals who occupy different social positions, and are therefore subject to different norms, values and expectations as well as different structural constraints. In the context of the hybrid kind model, we can

[1] I use "gender differences" in the theoretically naïve sense of everyday language to refer to observed differences between men and women, especially psychological and behavioural ones. I am not invoking the sex/gender distinction discussed in previous chapters.

interpret this as evolutionary psychologists being concerned with the base kinds, and social scientists with the status kind, of the hybrid kinds *men* and *women*.

A brief glance at the literature suggests an irreconcilable conflict between proponents of these two explanatory approaches. Evolutionary psychologists have developed an impressive amount of research which aims to show that psychological or behavioural differences between men and women are the product of evolved psychological adaptations. Feminist theorists criticise this research for being "genetic determinist" or "gender essentialist". These catchwords generally express the objection that someone is ignoring the environmental determinants of gender differences while naturalising and thus reinforcing the oppression of women. Feminists accordingly argue that we need to pay more attention to environmental factors, in particular that we should seek social explanations of gender differences.

Evolutionary psychologists reject feminist objections on the grounds that they are misrepresenting their science. They argue that, contrary to the allegations, evolutionary psychological explanations do account for a variety of environmental factors. Some proponents of evolutionary psychology go further than that. They insist that evolutionary psychological and social explanations of gender differences are compatible or even converging on similar conclusions. In either case, proponents of evolutionary psychological explanations suggest that the alleged tension with social explanations is merely based on a misunderstanding of the evolutionary facts.

Although the debate is quite complex, the advantages of choosing it as a case study are many. To begin with, there is an exceptionally rich amount of literature to base our inquiry upon. In addition, causal and political questions in this discussion are thoroughly intertwined, illustrating a further layer of complexity in investigating hybrid kinds. Most importantly, despite the huge amount of (often heated) discussion that the topic has inspired, a detailed account of how exactly the different causal as well as political claims relate is still outstanding. I believe that both positions on the relationship between evolutionary psychological and social explanations of gender differences are mistaken. More precisely, I think that the interpretations offered by both sides are inadequate in two related respects: (i) they fail to correctly locate causal-explanatory conflict between the two approaches and (ii) they do not provide an account of how exactly the different causal claims relate to political concerns in this discussion. As a result, the following inquiry not only serves as a case study for illustrating the potential challenges to causally explaining

hybrid kind phenomena. By providing a long overdue clarification of the debate on gender differences, it also makes an important contribution to the debate itself.

Section 5.1 will set the stage by pointing out a puzzle in the extant debate. While evolutionary psychologists can successfully fend off the objection that they neglect environmental factors, their rejoinder leaves us wondering if and where exactly the two parties disagree. In order to pin down this disagreement, it is prerequisite to understand what exactly the explanations proposed by each party are claiming. Section 5.2 tries to pinpoint the claims made by evolutionary psychologists by looking at their central concept of domain-specific psychological mechanisms. The concept runs together three ideas that need to be distinguished: domain-specificity, trigger innateness, and purpose-specific adaptation. I argue that only two of these, trigger innateness and purpose-specific adaptation, are relevant in the context of this debate. Section 5.3 turns to social explanations of gender differences. I propose a distinction between socialisation explanations and social structural explanations, as well as a distinction between three different types of socialisation explanations. Although the mechanisms picked out by these explanations tend to be empirically entwined, I argue that they need to be conceptually distinguished because they rely on different types of evidence. Section 5.4 puts together the results of the previous two sections. It examines how exactly the different types of evolutionary psychological explanations relate to different types of social explanations and explores implications for the debate on gender differences.

5.1 THE EXTANT DEBATE

5.1.1 Feminist objections to evolutionary psychology

As indicated above, feminist theorists tend to reject evolutionary psychological explanations for providing an "essentialist" or "genetic determinist" view of gender differences. Evolutionary psychological explanations, they argue, overestimate the relative importance of genes and understate the relative importance of environments in explaining human behaviour. Laurette Liesen, for instance, suggests that evolutionary psychologists "[…] downplay the flexibility of humans to respond to their current environments and circumstances" and therefore portray human behaviour as "extremely slow to change" (Liesen 2007, 52). In addition, feminists often object that evolutionary psychology justifies inequalities between men and women by naturalising them (Fausto-Sterling 2000; Contratto 2002). Hence, Betsy Lucal attributes her scepticism of evolutionary explanations of human behaviour to her impression that curiously many of these

explanations support current patterns of dominance (Lucal 2010, 47). Susan Contratto argues, more forcefully, that evolutionary psychology is "profoundly conservative" and "dangerous to feminism because it is used to justify and maintain the status quo" (Contratto 2002, 41). She concludes that the evolutionary psychological approach is "antithetical to change" in theory and practice and therefore has "no common ground" with feminist psychology (Contratto 2002, 43).

These remarks illustrate two distinct but related objections that feminists level against evolutionary psychologists: a causal-explanatory and a political (or normative) one. The causal-explanatory objection states that evolutionary psychological explanations of gender differences are incompatible with social explanations because the former deny or underestimate the role of socio-environmental determinants of human behaviour. The political objection states that evolutionary psychological explanations are detrimental to furthering gender equality because they justify the status quo and undermine political change. Supposedly – although this is surprisingly rarely made explicit – what connects these two objections is the assumption that evolutionary psychologists use explanatory falsehoods to support conservative politics. Much speaks in favour of this understanding. The burgeoning literature offering comprehensive methodological criticisms of the branch of evolutionary psychology most vocal in this debate certainly suggests that the claims should be taken with a pinch of salt.[2] However, I believe that matters are more complicated than extant appraisals of the debate acknowledge. This is true not only regarding the relationship between the causal claims made on each side, but also for the relationship between evolutionary psychologists' causal claims and feminists' political objections. In this chapter, I will focus on where exactly feminist social scientific approaches and evolutionary psychology disagree on the causal-explanatory level. Feminists' political demands and their relationship to the causal claims made by evolutionary psychology will be the topic of Chapter 6.

5.1.2 Evolutionary psychologists' defence against the causal-explanatory objection

Evolutionary psychologists have vehemently defended themselves against the objection that their approach is negligent of environmental determinants of behaviour. To the contrary, they argue, evolutionary psychology acknowledges a variety of ways in which socio-environmental factors can be relevant to human behaviour. In general, we can

[2] See, for instance, Gould & Lewontin 1979; Lloyd 1988; Buller 2005.

identify three ways in which evolutionary psychologists claim to account for the plasticity of human behaviour and psychology.

5.1.2.1 Environmental plasticity

The first type of plasticity recognised by evolutionary psychologists is *environmental plasticity*. Environmental plasticity is the idea that humans with a set psychological architecture are able to vary their behaviour in response to different environmental stimuli. For instance, Gangestad and Simpson (2000) suggest that humans switch between different mating strategies depending on the harshness of the environment. Because harsh environments require biparental care for reproductive success, so they argue, we should expect higher levels of monogamous mateship in more demanding environments and more promiscuous mateship in less demanding environments. In other words, environmental plasticity suggests that human behaviour and psychology can vary in response to different environments.

Importantly, although this form of plasticity accounts for some observed variation, it assumes that this variation runs along paths that have been trodden by natural selection. Environmental plasticity assumes that the reason an individual shows behaviour B_1 in environment E_1 and behaviour B_2 in environment E_2 is because of a history of evolutionary selection. According to this idea, the reason the individual varies his or her behaviour as described is because, in the evolutionary past of the individual's lineage, B_1 behaviour was adaptive in E_1-like environments, and B_2 behaviour was adaptive in E2-like environments. This idea is sometimes expressed in the claim that human possess largely identical *adaptive conditional strategies* which are "programmed" to respond differently to different environmental cues (Waynforth & Dunbar 1995; Thornhill & Palmer 2000; McKibbin et al. 2008).

5.1.2.2 Developmental plasticity

Evolutionary psychologists commonly distinguish environmental plasticity from *developmental plasticity*. Developmental plasticity is the idea that human psychological development can take different adaptive pathways depending on the cues in one's early developmental environment. This suggests that humans do not merely vary in behaviour because they use their existing psychological make-up in response to different environmental cues. Humans can also vary in behaviour as a result of different developmental trajectories, which lead them to develop diverse psychological make-ups (such as different dispositions or preferences) in the first place. Buss' explanation of the correlation between father-absence during childhood and incidence of casual sex in

women uses the concept of developmental plasticity. According to Buss, the correlation can be explained by the fact that father-absence is a developmental cue for polygamous sexual preferences. He suggests that the absence of fathers leads women to "conclude that men are not reliable investors", which makes them "pursue a strategy of extracting immediate resources from a number of short-term partners, rather than trying to secure the continued investment of one" (Buss 2003, 93). In other words, Buss argues that the observed correlation can be explained as an adaptive sexual strategy.

Note that, just like environmental plasticity, developmental plasticity is assumed to operate within the constraints set by our evolutionary past. Evolutionary psychologists claim that the reason an individual develops specific psychological characteristics in reaction to a certain developmental cue is because the developmental pathway in question conveyed a relative fitness advantage in past environments. In other words, both environmental plasticity and developmental plasticity follow paths established in the evolutionary pasts. The human mind in all its variability, according to this picture, is essentially a universal, two-level conditional programme shaped by natural selection.

5.1.2.3 Cross-cultural variation

In contrast to the previous to forms of evolutionary-shaped plasticity, evolutionary psychologists sometimes recognize a third type of plasticity – *cross-cultural variation* – that might lead humans astray of adaptive pathways. Cross-cultural variation is usually invoked to explain the fact that human psychology and behaviour varies across cultures. David Schmitt's discussion of "sociosexuality" (the preference for sex with a variety of different partners) is a case in point. Schmitt suggests that there "[…] may be certain aspects of culture that influence our evolved psychology in ways that accentuate or attenuate sex differences in sociosexuality" (Schmitt 2005, 252). Importantly, Schmitt acknowledges that, in contrast to environmental and developmental plasticity, these cultural influences may alter behaviour in not necessarily evolutionarily adaptive ways. As an example, Schmitt discusses the influence of "egalitarian sexual standards and gender role beliefs" which may partly explain the cross-cultural pattern of differences in sociosexuality (Schmitt 2005, 272). In other words, evolutionary psychologists acknowledge that social or environmental factors can sometimes affect human behaviour and psychology in ways that cannot be explained in terms of adaptive conditional strategies. However, this does not mean that they give up on the idea of a universal, evolved human cognitive architecture. Schmitt is careful to point out that cultural influences merely "accentuate" or "attenuate" evolved behavioural dispositions. Human culture, on this view, is not a

factor that freely gives shape to large parts of human cognitive architecture in the first place.

These are the three types of human plasticity evolutionary psychologists frequently invoke to respond to the objection that they are neglecting environmental determinants of behaviour. Are these concessions to plasticity enough to defuse the worries, and maybe even show that social and evolutionary psychological explanations of gender differences are compatible after all? Evolutionary psychologists Buss and Schmitt seem to think so. They claim that, contrary to feminist critiques, "feminists and evolutionary psychologists appear to *converge* on conceptualizations of human behavior as flexible and context-contingent" (Buss & Schmitt 2011, 771, my emphasis). This description of the relationship between social and evolutionary psychological explanations is quite contrary to the feminist portrayal discussed above. Instead of suggesting irreconcilable conflict over the importance of socio-environmental factors, evolutionary psychologists portray the relationship with social explanations as one of compatibility and even convergence on similar understandings of human behaviour. Before we disturb this idyllic image with further scrutiny, we need to consider a fourth conceptual tool that plays a crucial role in evolutionary psychologists' defence against feminist criticism.

5.1.2.4 The distinction between proximate and ultimate explanations

In addition to demonstrating their recognition of environmental determinants of human behaviour, evolutionary psychologists often invoke a second line of defence. In order to show that evolutionary psychological explanations are perfectly compatible with social explanations of human behaviour, they refer to the distinction between proximate and ultimate explanations (see Mayr 1963; Tinbergen 1963). Proximate explanations refer to causes that affect someone during their lifetime, such as developmental and environmental factors and cultural influences. Ultimate explanations, by contrast, refer to causes that affected the individual's evolutionary history, most prominently factors leading to natural selection. Proximate and ultimate explanations, so the idea goes, address distinct causal questions. Proximate explanations answer questions about the developmental or environmental (including cultural) determinants of a certain trait. Ultimate explanations answer questions about the evolutionary origin of the trait, that is, questions as to why the trait has been passed on through the lineage.

According to some proponents of evolutionary psychology, the proximate-ultimate distinction allows us to resolve the alleged conflict between evolutionary and social explanations of human behaviour. Social explanations, they claim, are concerned with

how the social environment affects the development and expression of an individual's traits during that individual's lifetime, hence are *proximate explanations*. Evolutionary psychological explanations, by contrast, are concerned with why these traits have evolved in the history of the human species, hence are *ultimate explanations*. Note that, on this understanding, evolutionary psychological explanations are restricted to describing the evolutionary underpinnings and adaptive function of human behaviour and psychology. Since they do not entail any claims about the proximate or developmental causes of human behaviour, they are simply incapable of carrying implications that could come into conflict with social explanations.

These two types of explanation, it is sometimes suggested, are compatible simply in virtue of the fact that address distinct causal questions. Heidi Colleran and Ruth Mace make this point with regard to their research on human behavioural ecology (Colleran & Mace 2011). They argue that their research, which explores the selective advantages of certain behavioural variants over others, focusses on ultimate causes. For that reason, they suggest, explanations of the same behavioural variant that refer to proximate causes are not incompatible (Colleran & Mace 2011, 290). Colleran and Mace believe that this understanding can be generalised to evolutionary and social explanations more generally. They suggest that evolutionary and social approaches "[…] can be understood as asking different 'whys', and really do not have to be mutually exclusive" (Colleran & Mace 2011, 292).

Sometimes defendants of evolutionary explanations go further than that. They argue that social and evolutionary explanations are not merely *compatible*, in the sense that they do not come into conflict, but suggest that the two are *complementary*, hence that both are required for a "complete" understanding of the phenomenon in question. Aaron Goetz et al., for instance, argue that their sperm competition hypothesis, which addresses the ultimate causes of rape, is complementary with feminist social explanations that explain rape as a phenomenon of domination and control (Goetz et al. 2008). Along similar lines, Alfonso Troisi argues that proximate and ultimate explanations of gender differences in vulnerability to social stress are complementary, and that both need to be taken into account to understand the phenomenon (Troisi 2001).

Evolutionary psychologists are not the first to suggest that ultimate and proximate explanations are compatible or even complementary – this is a common assumption in evolutionary biology (Mayr 1963; Tinbergen 1963). Section 5.4 will take a closer look at what exactly this statement entails. But it is important to note that arguments of this sort

make a crucial assumption. They presume that evolutionary explanations are restricted to describing the evolutionary underpinnings and adaptive function of human behaviour and psychology and do not entail any claims about the proximate or developmental causes of human behaviour. Regarding the discussion at hand, this assumption states that evolutionary psychological explanations are simply incapable of carrying implications that could come into conflict with social explanations, and vice versa. In other words, the argument from the proximate-ultimate distinction claims that although both evolutionary psychological and social explanations might be needed for a "complete" explanation of a specific human behaviour, the two types of explanations are, in a way, irrelevant to each other. We will come back to this assumption later.

5.1.3 Summary

In this section, we saw that evolutionary psychology acknowledges a range of environmentally determined behavioural flexibility. Contrary to feminists' allegations, it thus looks like there is no straightforward sense in which evolutionary psychological explanations can be considered genetic determinist or ignorant of environmental factors. Reference to the proximate-ultimate distinction further suggested that any apparent conflict between social and evolutionary psychological explanations can be resolved by recognising that both explanations answer different types of questions. Accordingly, in place of the predominant image of irreconcilable conflict, some evolutionary psychologists have suggested a relationship of compatibility and complementarity.

While further investigation is required to decide whether this response can really defuse feminists' reservations about evolutionary psychology, one thing has become apparent. Contrary to what the heated discussion suggests, it is not at all clear where exactly proponents of social explanation and proponents of evolutionary psychological explanations disagree. If there is causal-explanatory disagreement between the two types of explanations, it does not simply consist in the fact that the former assert the relevance of social and environmental factors whereas the latter deny it. In order to decide whether social and evolutionary psychological explanations of gender differences really are incompatible, we first need to develop a better understanding of what each type of explanations is claiming. The next section makes a start on this task by looking at evolutionary psychology.

5.2 WHAT IS AN EVOLUTIONARY PSYCHOLOGICAL EXPLANATION?

In the previous section, we encountered a puzzle regarding the relationship between evolutionary psychological and social explanations. Whereas feminists mostly suggest that the two are incompatible, proponents of evolutionary psychology have been eager to portray a more harmonious relationship. In this section, I argue that the perceived difficulty in identifying whether there is actual causal-explanatory conflict underlying the discussion at hand is not merely imagined. It results from a genuine confusion about the content of one of evolutionary psychology's central concepts. Evolutionary psychologists who figure most prominently in the debate on gender differences generally emphasize that the human mind is a cluster of *domain-specific psychological mechanisms* (e.g. Buss 1990; Barkow 2006; Pinker 2002; Tooby & Cosmides 1992; Tooby & Cosmides 2005). A central claim of their explanations of gender differences is that the differences in question are the product of these domain-specific psychological mechanisms. In order to pin down the content of evolutionary psychological explanations, we therefore need to explore the concept of domain-specific mechanisms in more detail.

5.2.1 Domain-specificity and the blank slate

Evolutionary psychologists generally contrast domain-specific psychological mechanisms with *domain-general* psychological mechanisms. The concept of domain-general psychological mechanisms, in turn, is used to describe what evolutionary psychologists take to be the prevalent social scientific understanding of human psychology: the idea that a single or very few general learning mechanisms are responsible for most phenomena of human psychology and behaviour. According to evolutionary psychologists, the domain-general mechanisms assumed by the social sciences are "[…] equipotential, content-free, content-independent, general-purpose, domain-general" or "[…] constructed in such a way that they can absorb any kind of cultural message or environmental input equally well" (Tooby & Cosmides 1992, 29).

According to this view, evolution has equipped the human mind with little more than a general purpose mechanism for learning, and all human behaviour is the product of learning or culture. Evolutionary psychologists usually refer to this model as the *Blank Slate Model* or the *Standard Social Science Model* (*SSSM*) and they strongly disagree with it. They are at pains to point out that the SSSM is misguided and ought to be replaced with an understanding of the mind that is informed by evolutionary psychology (see Pinker

2002 for a book-length discussion). On these evolutionary psychologists' understanding, the human mind is composed of a variety of domain-specific psychological adaptations.

There are several reasons to be cautious about evolutionary psychologists' portrayal of the social sciences as "blank slatists". Firstly, the SSSM can most plausibly be ascribed only to very specific traditions in sociology and cultural anthropology – such as those associated with the work of Emile Durkheim, Margaret Mead, Franz Boas or Alfred Kroeber. Yet evolutionary psychologists tend to indiscriminately ascribe this model to social science as a whole. Hence the term "standard social science model". This sweeping generalisation is rarely backed up with any references from social scientific publications. If references are included (as in Pinker 2002) they tend to be focussed on specific social scientific traditions and hence fail to be representative of the social sciences more generally. Most of all, there is little of no engagement with contemporary feminist critics of evolutionary psychology, despite the fact that they are one of the main targets of the SSSM objection. In other words, while there might be "blank slatists" among the social sciences, evolutionary psychologists generally fail to demonstrate that this view is shared by their feminist opponents, let alone the social sciences as a whole.

In addition, evolutionary psychologists' criticism of the blank slate model generally fails to acknowledge the fact that the model can be understood in two ways – either in terms of an *ontological,* or in terms of an *explanatory* commitment. Evolutionary psychologists typically understand the model in terms of an ontological commitment. This ontological commitment states that the human mind is constituted by a domain-general mechanism for learning, and that human behaviour is therefore largely determined by socialisation and culture. Interpreted as an *explanatory* commitment, by contrast, the model says that although facts about human biology and evolution are causally relevant for human behaviour, human behaviour cannot be *understood* in terms facts about biology and evolution. In order to understand why humans act and think the way they do, we need to make reference to their culture and society.

This distinction allows us to recognise that at least some social scientists which have been accused of "blank slatism" may be not so much committed to the ontological claim that behaviour *is only caused* by culture and learning as to the explanatory claim that human behaviour *is best explained* in terms of culture and learning. When criticising social explanations of human behaviour, evolutionary psychologists typically base their assessment on the assumption that their opponents are committed to the ontological claim. They generally do not consider the possibility that the commitment could be merely

explanatory. In several cases, philosophers have claimed that this is wrong. The alleged "blank slatists", they argue, do not so much *deny* genetic, evolutionary, or biological causes of human behaviour as to firmly *ignore* them (Longino 1990; Longino 2012; Kronfeldner 2009; Jackson 2010). Moreover, they have insisted that this happens for sound methodological reasons. I will come back to this "right to ignore" in Chapter 6. For now, it suffices to note that social scientists who show evidence of using the SSSM might be committed to the explanatory claim rather than the ontological one. In this case, evolutionary psychologists would be misrepresenting the social scientist' position.

These two problems imply several challenges when it comes to making sense of the tensions between proponents of social and evolutionary psychological explanations of gender differences. Firstly, proponents of social explanations of gender differences might not be committed to the blank slate model at all. Nothing said so far rules out the possibility that one can coherently claim that a trait is *both* the product of learning or culture *and* the result of a specialized cognitive adaptation. Similarly, even if someone claims that a particular gender differentiated behaviour is the product of society or culture, this would not commit them to the idea that the same is true for all other characteristics of human behaviour or psychology. Since evolutionary psychologists have failed to show that proponents of social explanations are committed to the blank slate model, these options cannot be ruled out.

Secondly, the discussion above suggests that the dispute in question might not be rooted in actual disagreements about what happens on the causal-ontological level. Instead, proponents of social explanations may object to evolutionary psychological explanations on the basis of an explanatory commitment to the blank slate model. In this case, proponents of social explanations reject evolutionary psychological explanations not because they believe these are factually wrong, but because they believe they provide somehow bad, misleading or unhelpful explanations of gender differences. If this is correct, the disagreement would not be a disagreement about causal facts, but a disagreement about the requirements of good explanation, which may depend on pragmatic and context-specific considerations. The next chapter will look in more detail at disagreements of this type, in particular the potential for a "right to ignore" certain claims about causal facts irrespectively of whether they are true or false. For now, we will focus on (real or imagined) disagreements about causal facts.

Besides the two problems outlined above, there is a third and even weightier concern which brings us back to the concept of domain-specific psychological mechanisms. So

far, the concept is still lacking in clear, positive content. We have identified that the idea of domain-specific psychological mechanisms is at the heart of the debate over evolutionary psychological versus social explanations of gender differences. Nevertheless, it seems that the concept primarily serves the negative function of saying how a behaviour has *not* been brought about. As a result, it is not clear what characterises a behaviour as the product of domain-specific psychological mechanisms, other than the fact that it is not the product of *domain-general* learning mechanisms. In the following, we will see that this lack of detail obscures substantial ambiguity in evolutionary psychologists' claims.

5.2.2 Domain-specific mechanisms as purpose-specific adaptations

One possible interpretation of the idea of domain-specific mechanisms comes from evolutionary psychologists' commitment to ultimate explanations. If we take this commitment seriously, it suggests that domain-specific mechanisms are best understood as *purpose-specific psychological adaptations*. According to this understanding, to say that a behaviour is the product of a domain-specific psychological mechanism is to make a claim about the behaviour's evolutionary history. It amounts to saying that the reason individuals today display the behaviour is because the psychological mechanisms that produces it has been selected for producing the behaviour in question. In other words, the behaviour in question is the product of a purpose-specific adaptation which conveyed a fitness benefit on our ancestors.

In the context of the debate on gender differences, this understanding suggests that evolutionary psychologists' claims that gender differences are caused by domain-specific mechanisms effectively says something like the following: gender-differentiated traits are produced by psychological mechanisms which have been selected for producing the traits in question. This understanding assumes that the selective pressures on male and female psychology have been very different in certain regards. As a result of this, men and women today possess different purpose-specific psychological adaptations which are responsible for observed gender differences in psychology or behaviour.

Presumably, on this view, the opposing claim that gender differences are caused by domain-general mechanisms states that gender differences are the product of a purpose-general adaptation. A plausible candidate for such a purpose-general adaptation would be the human capacity for learning. This seems to suggest that saying gender differences are caused by domain-specific mechanism means that they are not the product of learning. An interpretation of domain-specificity along these lines therefore dovetails nicely with evolutionary psychologists' aversion to the idea that human behaviour is mainly the

product of learning. But there is a caveat here. Saying that a trait is the product of a purpose-specific adaptation mainly states that the trait persists in the lineage because it has been selected for a specific effect. Contemporary evolutionary theory, however, recognises several ways in which a domain-general adaptation like learning itself can be involved in bringing about more purpose-specific adaptations (Jablonka & Lamb 2005). There is reason to believe that bird songs, for instance, are purpose-specific adaptations that are transmitted down generations by learning. In the human realm, arguments to this effect have been made, for instance, with regard to tool use and imitation (see, for instance, Sterelny 2005; Heyes 2011; 2012).

On an adaptationist understanding of domain-specificity, then, gender differences brought about by learning could be considered the result of domain-specific mechanisms if the fact that men and women learn different traits is itself a sex-specific adaption. We could imagine, for instance, that children learn gender-differentiated mate preferences from their parents, and that this learning pattern spread in the population because it conveyed a relative fitness benefit. Hypotheses of this sort are rarely considered in any detail in the extant discussion on gender differences. I will come back to it in Section 5.5, after laying out the different interpretations of evolutionary psychological and social explanations in more detail. For now, we merely need to note that saying a gender-differentiated behaviour is the product of a purpose-specific adaptation does not necessarily rule out that the behaviour is learned.

Incidentally, the interpretation in terms of purpose-specific adaptations reflects how more charitable critics of evolutionary psychology tend to understand the notion of domain-specific mechanisms. They suggest that evolutionary psychology is defensible as long as proponents strictly adhere to their commitment to ultimate explanation. While evolutionary psychologists can tell us whether a specific trait is an adaptation, they should leave claims about proximate or developmental mechanisms to sciences who are in the business of empirically investigating them – such as developmental biology, psychology, or empirical social science (Lewens 2003; Goldfinch 2015). At times, evolutionary psychologists adhere to this recommendation. They insist that claims about ultimate causation do not entail any implications with respect to the proximate mechanism that reliably reproduce these mechanisms in each generation of individuals. Consider, for example, a study by David Buss which found cross-cultural evidence (from 33 countries from six continents) for differences in mate preferences between men and women (Buss 1989a). The findings suggests that, on average, women put a higher value on signs of

resource acquisition such as wealth and education than men. Men, by comparison, put a relatively higher value on signs of reproductive capacity, such as youthful appearance and beauty. According to Buss, these finding support the evolutionary (i.e. ultimate) hypothesis that men and women possess "adaptations to sex-differentiated reproductive constraints in our evolutionary past" (Buss 1989a, 14). To clarify that he is only committed to the ultimate hypothesis, Buss is cautious to insist that

> […] these results yield little information about the proximate (social, psychological, physiological, ontogenetic) mechanisms directly responsible for their existence. Possible candidates include genetic differences between the sexes, sensory preferences analogous to food preferences, socialization differences during development, and structural effects at a societal level such as those that limit female access to economic resources […] research on proximate mechanisms is needed to develop a more complete explanatory account of observed sex differences in mate preferences.
>
> (Buss 1989a, 13)

In this passage, Buss emphasises that his findings only relate to the evolutionary (or ultimate) hypothesis that the observed partner preferences are purpose-specific adaptations for mate choice. Questions about the proximate causes of these adaptive preferences, by contrast, are still open and might turn out to involve socialisation or even social structural features. Importantly, Buss' discussion recognises the caveat noted above. He acknowledges that the (ultimate causation) hypothesis that a behaviour is the product of a specialised cognitive adaption is compatible with the (proximate causation) hypothesis that the behaviour is due to learning or even social structural causes.

More often, however, evolutionary psychologists' characterisations of domain-specific mechanisms straddle proximate and ultimate claims – in spite of their own commitment to carefully distinguishing between the two. Evolutionary psychologists tend to overstep the boundary from ultimate to proximate explanations by attacking social explanations of the same behaviour. They suggest that the claim that a behaviour is the product of domain-specific mechanisms implies that it *cannot* be attributed to socialisation or social learning – both of which are evidently proximate mechanisms. Donald Symons, for instance, disagrees with Buss' above statement and argues that he should be more decisive about the implications for proximate mechanisms (Symons 1989). In particular, Symons suggests that social explanations of gender differences are incompatible with Buss' hypothesis of domain-specific mechanisms for mate preferences:

> [social] scientists typically attribute human mate preferences, and sex differences therein, to such things as "cultural conditioning", "socialization", "social learning" and "stereotyped sex roles"[…] all of which imply that these preferences are underpinned by some sort of

> generalized brain/mind mechanism (presumably of association or symbol manipulation); in other words, such theories imply that *specialized mechanisms of mate preference do not exist*.
>
> (Symons 1989, 34, my emphasis)

In other words, Symons seems to believe that a behaviour cannot both be the result of a domain-specific mechanisms and have social factors as proximate causes. We could cast off Symons' objection by insisting that evolutionary psychology's domain-specific mechanism simply are purpose-specific adaptations, and that Symons must be confusing his terminology. Properly understood, Buss' hypothesis that gender differences are due to domain-specific mechanisms is compatible with the proximate mechanisms invoked in social explanations. But this would ignore the fact that Symons' line of argument is representative of how many evolutionary psychologists seem to think about domain-specific mechanisms. In the context of the debate on gender differences, it is fairly common for evolutionary psychologists to say, for instance, that traits which are due to domain-specific mechanism cannot be the product of learning. This suggests that Buss is invoking an alternative notion of domain-specific mechanism, according to which such mechanisms entail information about proximate causation after all. In the following section, I explore what exactly that notion is.

5.2.3 Domain-specificity and innateness

To understand what evolutionary psychologists' mean when they talk about domain-specific mechanisms in the way described above, it is helpful to consider Muhammad Khalidi's analyses of the concepts of domain-specificity and innateness (Khalidi 2010; Khalidi 2001). To avoid terminological confusion, begin with Khalidi's account of domain-specificity. For Khalidi, a domain-specific psychological mechanism (or "cognitive system", in his words) is a mechanism that is in principle generalizable to new domains but fails to do so in practice for reasons having to do with the system's evolutionary history. In Khalidi's words, a domain-specific cognitive system is a system that "[…] systematically fails to yield a correct result in the case of stimuli that the system did not evolve to deal with" (Khalidi 2010, 196). As an example, he discusses the alarm calls of vervet monkeys. Researchers found that vervet monkeys utter three distinct alarm calls to warn their group members from three types of predators (leopards, eagles, and snakes). Yet they fail to produce distinct alarms calls for other types of predators that are prevalent in their environment. According to Khalidi, the mechanisms that underlies the alarm calls is domain-specific because it could in principle be applied to produce alarm calls for other predators that are frequently encountered, yet fails to do so in practice.

The above analysis offers a notion of domain-specificity that entails information about proximate mechanisms. It tells us that domain-specific mechanisms have specific information-processing characteristics in their role as proximate causes. But there is reason to doubt that it is the notion of domain-specificity used by evolutionary psychologists. While Khalidi claims that his analysis elucidates what domain-specificity means in the cognitive sciences more generally, he suggests that evolutionary psychologists are guilty of misusing this concept (Khalidi 2010, 195). To illustrate this, he points to a crucial mismatch between evolutionary psychologists' use of the concept of domain-specificity and his own analysis. Evolutionary psychologists regularly use the concept of domain-specificity to describe preferences (such as preferences for particular mate characteristics, or for short versus long-term mating). But preferences are not the kind of entities that could be domain-specific in Khalidi's sense. This, Khalidi argues, is because preferences are usually not in-principle generalizable. Instead of relying on a rule or principle that could in principle be applied to other domains but fails to do so in practice, the proposed mechanisms for mate selection rely on information of a specific subject matter – mating and reproduction (Khalidi 2010, 195-196). As a result, preferences for physical features of potential mates are not in principle generalizable to other domains such as food choice, because both domains involve entirely different sets of stimuli.

Since evolutionary psychologists frequently use the term "domain-specificity" to refer to preferences and other traits which are not in-principle generalizable, it would be fair to suspect that they may have something entirely different in mind when using the term. This idea gains tractions once we consider common beliefs about the relationship between domain-specificity and innateness. In a second paper, Khalidi observes that many people (including many evolutionary psychologists) assume that the two concepts are intrinsically connected (Khalidi 2001). They believe that traits which are innate are the product of domain-specific mechanisms, and vice versa (Khalidi 2001, 192). Khalidi, however, suggests that this is wrong. The two concepts are independent – a trait can be the product of domain-specific mechanisms without being innate, and vice versa. This opens up an interesting possibility. When talking about "domain-specific" mechanism as proximate causes, evolutionary psychologists might actually be referring to a very different concept – the concept of innateness.

Incidentally, some evolutionary psychologists used to describe domain-specific mechanisms as "innate" in earlier writings (see, for instance, Tooby & Cosmides 1989a; 1989b; 1990), but seem to have renounced this terminology in more recent publications.

Instead, newer publication talk about the innate only in scare quotes, and often for the purpose of illustrating social scientists' misunderstanding of evolutionary psychology (Tooby & Cosmides 1992; Tooby & Cosmides 2005). This shift in terminology might have been motivated just as much by political controversy as by conceptual disputes. The concept of innateness has been subject to a huge amount of conceptual scrutiny. Some people argue that the concept is hopelessly confused and is best done away with in the scientific literature (Bateson 1991; Bateson & Martin 1999; Griffiths & Machery 2008). Other suggest that it is best understood along the lines of environmental canalization (Ariew 1996; Ariew 1999).

Khalidi's argument, however, relies on a different understanding of innateness – *trigger* (or *dispositional*) *innateness*. The notion of trigger innateness has been first introduced by Stephen Stich (Stich 1975). It suggests that a psychological trait or capacity is innate if it is susceptible to a *poverty of the stimulus* argument. The paradigm example of a poverty of the stimulus argument is Noam Chomsky's argument about innate grammar in human language (Chomsky 1957; Chomsky 1966). According to Chomsky, children could not possibly acquire the complex rules of grammar on the basis of the language information they are exposed to alone. The reason this is impossible, he suggests, is that they are not exposed to comprehensive information about which sentence structures are not grammatical. Nevertheless, children come to know the grammar of their native language – they know which sentences are grammatical and which are not. Chomsky concludes that the informational deficit, i.e. the "missing" information about which sentence constructions are ungrammatical, has to be innate.

Against this background, Khalidi suggests that a human cognitive feature is innate if there is an "informational deficit" between the learning input an individual receives, and the behavioural output that it produces (Khalidi 2001, 193). He concludes that once we understand the concepts of innateness and domain-specificity in these ways, we can see that they are independent. In particular, contrary to what evolutionary psychologists seem to assume, domain-specific capacities can be the product of learning, and domain-general capacities can be innate (Khalidi 2001, 196). However, evolutionary psychologists could object to this by arguing a psychological capacity that is domain-specific can achieve this resistance to generalisation only in virtue of an innate component. If the psychological capacity was entirely the product of general learning mechanisms, so they could argue, we should expect it to extend to a new domain easily via learning, because general learning mechanisms are unable to discriminate between different domains.

To see that this is not the case, consider again the example of vervet monkeys. In principle, we need not make recourse to an innate component in order to explain why the very same learning mechanisms leads to the development of predator-specific calls in some domains (leopards, eagles, snakes) but not others (humans, diggers, dogs). Instead, the ability to discriminate between domains could be the product of learning. For instance, parent monkeys could condition their offspring early on to pay particular interest to some predators (i.e. leopards, eagles, snakes) but not others. As a result of this conditioning, the monkeys are "primed" to develop specific alarm calls in reaction to these predators but not others. In other words, domain discrimination could be achieved not by innate components, but by a specific learning trajectory which primes individuals towards certain domains. Although I am not aware of any research suggesting that this is the actual developmental trajectory for monkey calls, priming by learning is an abundant mechanism in human psychology. Trained musicians tend to be more alert to musical patterns than others, trained philosophers tend to be more perceptive of the logical structure of arguments, etc. If this reply is correct, Khalidi's argument for the independence of domain-specificity and innateness stands.

5.2.4 Domain-specificity, innateness, and gender differences

The above discussion suggests that evolutionary psychologists who want to infer claims about proximate causation from the concept of domain-specific mechanisms are not talking about domain-specificity at all. Instead, the phenomenon they intend to refer to is trigger innateness. There are several additional considerations that support this idea. For one thing, the notion of trigger innateness provides a neat contrast between "innate" and "learned" that suits the argumentative strategy used by many evolutionary psychologists. To say that a capacity is trigger innate is to say that *not* all the information necessary to perform the capacity has been acquired by learning – some of it must come from "within" the individual, for instance in the form of genetic representation (Shea 2012). This dovetails nicely with evolutionary psychologists' engagement of environmental factors as detailed in Section 5.1. There, we could see that evolutionary psychologists tend to think of environmental factors as "cues" for adaptive conditional strategies. This suggests that many evolutionary psychologists think of environmental factors primarily as triggers for behavioural responses that have been "hard-wired" in our evolutionary past. For them, the concept of trigger innateness provides a coherent way of spelling out the claim that gender differences are not merely the product of learning while acknowledging the (triggering) influence of environmental factors.

Note that we cannot make sense of this characteristic if we interpret evolutionary psychologists' claim in terms of domain-specificity or purpose-specific adaptation. Neither concept provides a coherent interpretation of the claim that gender differences are "not merely the product of learning". The claim that gender-differences are due to domain-specific mechanisms, in the sense detailed above, does not rule out the possibility that the mechanisms have been domain-specialised through a particular learning trajectory. At the same time, claiming that gender differences are purpose-specific adaptation does not imply that they cannot be learned. Learned traits, as argued above, can be purpose-specific adaptations.

In addition, interpreting evolutionary psychologists' claims in terms of trigger innateness enables us to make sense of the political tension in the debate in a way the notions of domain-specificity or purpose-specific adaption do not. Neither purpose-specific adaptation nor domain-specificity are of much interest in the debate over gender differences. Nothing in the literature suggests that proponents of social explanations take any issue with the suggestion that psychological mechanisms responsible for gender differences fail to generalise to other domains. And while some people take issue with the suggestion that gender differences are adaptations, this is usually down to the fact that they are unaware of the fact that adaptations can be transmitted by learning. Instead, as suggested above, they typically associate such claims with the suggestion that the differences in question are *innate*.

In contrast to domain-specificity and purpose-specific adaptations, the concept of innateness has been the direct target of endless political controversies (Block & Dworkin 1976; Lewontin et al. 1984; Fausto-Sterling 1992; Segerstråle 2000). These controversies are generally based on the assumption that innate traits are immutable or otherwise outside our responsibility. Given the ambiguity and confusion surrounding the concept of innateness, it is not possible to straightforwardly decide whether this assumption is justified. But this does not detract from the fact that the assumption exists and continues to fuel political controversy. This suggests that the widespread feminist opposition to evolutionary psychological explanations of gender differences may reflect feminists picking up on the fact that such explanations continue to invoke ideas of innateness. Rather than struggling to pin down claims about domain-specific psychological mechanisms, they may have recognised them for what they are– claims about innateness in disguise.

5.2.5 Summary

The discussion above suggests there are two ways of interpreting evolutionary psychologists' reference to "domain-specific" psychological mechanisms, neither of which involves domain-specificity as understood in the cognitive sciences. The first interpretation takes seriously the commitment of some evolutionary psychologists to only make claims about proximate causation. On this interpretation, domain-specific psychological mechanisms are purpose-specific adaptations. Evolutionary psychological explanations, on this view, state that particular gender differences are adaptations, but without making claims about the proximate causes for these differences.

On the second interpretation, and despite evolutionary psychologists' attempts to eliminate the "innate" from their writings, their explanations are best understood as making claims about a specific form of innateness called trigger innateness. Trigger innateness not only provides an understanding of the innate precise enough to be of scientific use, but also allows us to make sense of two common characteristics of evolutionary psychological explanations: the often suggested contrast with "pure learning" explanations, and their immediate political explosiveness. However, insofar as evolutionary psychologists make claims about trigger innateness, they betray their commitment to restricting themselves to ultimate explanations. Together with this commitment, the idea vanishes that social and evolutionary psychological explanations are compatible simply in virtue of the fact that they address different causal questions. If both social and evolutionary psychological explanations address proximate factors of gender differences, we need to take a much closer look to find out whether these explanations are compatible. But before we can address this question, we first need to clarify the nature of social explanations.

5.3 WHAT IS A SOCIAL EXPLANATION?

5.3.1 Socialisation and social structures

An obvious way to think about social explanations, and the one favoured by many evolutionary psychologists, is in terms of *learning*. According to this idea, social explanations understand human psychology and behaviour as shaped by learning processes in a social context. In the terminology of sociologists, this form of learning is referred to as *socialisation*. Socialisation comprises a diverse range of learning processes by which an individual comes to understand, predict and to varying degrees internalise and adhere to the norms, values, expectations and traditions of the society they live in.

In the context of the debate at hand, the focus lies specifically on *gender socialisation*, that is, social learning processes which lead individuals to acquire gender-differentiated psychological characteristics or behavioural dispositions. An example of a socialisation explanation of gender differences would be the following. Assume it's true that men, on average, place a higher value than women on a potential partner's youth and physical attractiveness. A social explanation in terms of gender socialisation could argue that this is because men and women are taught very different values regarding partner choice. For instance, boys and men tend to be more exposed to the idea, be it via the media, peers, or parents, that having a beautiful partner is an important symbol of personal success and social status. Girls and women, by contrast, are more frequently confronted with the message that partners should be physically or economically powerful. These differences in socialisation, so the argument continues, lead men more than women to internalise the idea that youthfulness and beauty are very desirable characteristics in a partner. As a result, men explicitly or implicitly use these criteria in partner choice more so than women do.

There are several different pathways in which socialisation can influence human behaviour and psychology. As the example above suggests, socialisation explanations often involve *internalisation*. Internalisation, here, is understood as the process by which individuals make a social norm or expectation part of their own system of beliefs and values. One way in which socialisation can influence people's behaviour is by *conscious internalisation*. Individuals who consciously internalise certain norms are aware that they support the norms in question. Unless they see reason to conceal their views, these individuals will typically express support for the norms in targeted questionnaires and interviews. *Subconscious internalisation*, by contrast, cannot be picked up so easily. Subconscious internalisation occurs when social norms, values or beliefs influence an individual's psychology and behaviour without being reflected in the individual's explicitly held values and beliefs. This phenomenon is usually discussed in the literature in the context of implicit bias. An increasing number of studies on implicit bias show that individuals can discriminate against individuals in accordance with gender or racial stereotypes even when their explicit convictions are egalitarian (Greenwald & Banaji 1995; Greenwald & Krieger 2006). This suggests that socialisation may occur subconsciously, without affecting an individual's explicit attitudes.

In addition to this, socialisation need not involve internalisation at all – what is socially learned need not become part of one's personal belief and value systems. Instead, individuals might simply adapt their preferences and behaviour to social norms, values,

or traditions in an instrumental manner. If individuals adapt a social norm instrumentally, they act in accordance with the norm not because they believe it to be right or good, but simply because they know that the norm is socially enforced and that failure to comply will in some way be punished. As an example, we could imagine a case where a man prefers young and physically attractive women not because he believes they are intrinsically more desirable partners, but because he knows his peers will think less of him if his partner does not fulfil this description. Both cases – subconscious internalisation and instrumental adaption of social norms – suggest that the influence of socialisation cannot always be picked up straightforwardly. In particular, it is unlikely to show up in people's reports of their own beliefs and attitudes and therefore requires additional methodological tools.

Although this discussion is not meant to be exhaustive, it gives us a good idea of the diverse pathways by which socialisation can affect human behaviour and psychology. However, there is another form of social explanation which does not rely on the idea of socialisation at all. These are so-called *social structural explanations*. Social structural explanations have been discussed in detail by Sally Haslanger (Haslanger 2016; Haslanger 2015; see also Jackson & Pettit 1992). According to Haslanger, social structural explanations explain the behaviour of an individual by referring to the position the individual occupies in a social structure without having to invoke individual psychological differences. Social structures, in turn, are networks of social relations which can obtain either between different individuals or between individuals and things. Importantly, social structures can constrain the behaviour of individuals in ways which can be explanatorily relevant.

As an example, Haslanger discusses a social structural explanation for the continuing economic disadvantage of women relative to men (Haslanger 2016; see also Okin 1989; Cudd 2006). The example is illustrated with a couple, Lisa and Larry, who decide to have children. Lisa and Larry are completely identical with respect to their education, talents, interests, and intelligence. As a result, they are both perfectly equally qualified when it comes to either paid employment or child care/domestic work. However, given the society they live in, Lisa and Larry a facing a set of structural constraints when making the decision to have a child. The relevant constraints include having to do paid work to support their family, a shortage in affordable child care, and, importantly, a wage gap between men and women with women, on average, earning twenty-five percent less than men. While the need for paid work and the shortage in childcare in principle affect Lisa

and Larry equally, the gender wage gap affects them differently *qua their social position* as man/woman. This has a crucial impact on Lisa's and Larry's likely arrangements for childcare. According to Haslanger, the difference in earnings makes it "[…] reasonable for Larry to work fulltime and for Lisa to make adjustments in her work, e.g., to work part-time, to take time off, to take a less demanding job" (Haslanger 2016). Furthermore, insofar as most heterosexual couples in a society are subject to the same constraints as Lisa and Larry, their reasoning is representative of a general pattern in that society. Even if partners in all heterosexual couples are psychologically identical and perfectly rational, given the structural constraints, they will arrive at the same arrangements for distributing paid and domestic work.

In addition, the upshot of this structurally caused collective decision pattern is to *reinforce* the structures that gave rise to it. For not only will it leave the majority of women with less economic power in their relationships, it also reinforces the existing wage gap between men and women. After all, from the perspective of employers, the decision pattern suggests that women are more likely than men to leave their career or to split their energies between paid work and housework. This makes women more risky job candidates than men. As a result, women are likely to be considered eligible only for jobs that require less commitment, mobility, or experience, and are therefore less financially rewarded. Thus, women's relative economic disadvantage has been reinforced.

5.3.2 The relationship between different types of social explanations

In this section, we will take a brief look at how the different types of social explanations relate and to what extent evolutionary psychologists are aware of this. From what we have learned so far, we know that evolutionary psychologists think of social explanations primarily as socialisation explanations that make reference to learned differences. But there are important exceptions. David Buss and Michael Barnes, for instance, distinguish "socialisation" and "structural effects" (Buss & Barnes 1986). As an example of a social explanation that invokes structural effects, they discuss the "female economic powerlessness" hypothesis (Buss & Barnes 1986; see also Buss 1989a). The female economic powerlessness hypothesis is meant to explain women's preference for partners with signs of high earning capacity, as measured in self-report questionnaires. The hypothesis suggests that women's preference can be explained by the fact that women have less access to economic resources and power than men, and can therefore best attain these resources by choosing a mate that can provide them (Buss & Barnes 1986, 569).

Despite distinguishing socialisation and social structural processes conceptually, Buss and Barnes later suggest that the two are empirically connected to a degree that, for them, seems to make it unnecessary to discuss each as a causal mechanism in its own right. They claim that "traditional socialization practices are presumed to maintain and support these structural differences, and are used to inculcate role-appropriate values in males and females" (Buss & Barnes 1986. 569). Consequently, in all of the following discussion, socialisation and social structural explanations are treated as a single "Structural Powerlessness and Sex Role Socialization" hypothesis for the gender differences in mate preferences (Buss 1989a; Buss 1989b; Symons 1989)

The claim that socialisation and social structural mechanisms are empirically connected is certainly true to an extent. Socialisation and social structures are intricately related and in many cases mutually constitutive. After all, individuals are generally subject to particular types of socialisation (e.g. gender socialisation, professional socialisation) *qua* occupying a certain social position. Often, socialisation pressures on these positions will work to stabilise and reinforce the social structure. Moreover, it seems that social structural features themselves could give rise to socialisation processes. In our examples, the fact that Lisa and Larry confront different social constraints might lead them to develop desires, preferences, and expectations. For instance, Larry's desire for a professional career might plausibly be reinforced by the relative prospects of success. By contrast, in light of her professional challenges, Lisa might learn to "not want what she can't have" and attribute higher value to things like spending time with her child and looking after the house. In this case, differential structural constraints made Larry and Lisa develop different psychological characteristics, hence set off a process of gender-differentiated socialisation in its own right.

Although socialisation and social structures are likely to be intertwined in the way Buss and Barnes suggest, the problems with entirely collapsing the different aspects into a single hypothesis are several. First of all, Buss and Barnes' characterisation suggest that socialisation and social structural features always point in the same direction. This ignores the possibility that socialisation and social structures can have a significant degree of independence. Socialisation processes, for instance, need not support existing social structures unequivocally. Imagine that a government decides to address the gender wage gap by implementing a policy for progressive gender role socialisation that aims to get more women into the better-paid traditionally male professions. If successful, this "socialisation reform" would counter rather than reinforce the social structures that

disadvantage women economically. Now imagine the scenario of a society that managed to abolish the wage gap and its resulting social structural effects. In such a scenario, residuals of traditional gender role socialisation could lead women to continue doing the majority of housework despite equal or higher professional success than their male partners. These examples suggest that socialisation can precede as well as lag behind social structural change – the influence of socialisation and social structures need not be consistent.

The same is true the other way around. Social structures can effect gender-differentiated behaviour without requiring any supporting socialisation, that is, without requiring learned (or, for that matter, innate) differences in men's and women's underlying psychology. Consider again the example of Lisa and Larry. The example presumed that Lisa and Larry are entirely *identical* with regard to their education, talents, interests, and intelligence, yet settle on a division of labour in accordance with traditional gender stereotypes. This suggests that social structures like the wage gap may lead to a gender-differentiated pattern of behaviour even when men and women have not been exposed to any gender-differentiating socialisation. In other words, while socialisation explanations explain gender differences in behaviour by referring to learned differences in men's and women's psychology, social structural explanations do not require psychological differences of any kind. Instead, they can explain such differences simply by pointing to the differential structural constraints faced by men and women.

In addition to conflating socialisation and social structural explanations, Buss and Barnes' fail to acknowledge different types of socialisation explanations. Their claim that socialisation processes "inculcate role-appropriate values in males and females" suggests that gender role socialisation always involves internalisation of the relevant values. This description fails to acknowledge the difference between conscious and subconscious internalisation, which – as argued above – require different methods for being picked up empirically. It also fails to acknowledge that some socialisation processes do not involve internalisation at all. As discussed above, individuals might decide to adhere to social norms for purely instrumental reasons, without making the norm a conscious or subconscious part of their belief system.

In other words, when discussing social explanations of gender differences, evolutionary psychologists do not always acknowledge the different types of social explanations and the underlying causal pathways that they represent. This omission may have a grave impact on evolutionary psychologists' scientific reasoning. Distinguishing

the different social processes is not simply a matter of acknowledging the complexity of social mechanisms for gender differences. It has important implications because each of the mechanisms described above potentially requires different kinds of evidence to be supported or dismissed.

This problem becomes apparent in Schmitt's discussion of gender differences in sociosexuality introduced above (Schmitt 2005). Schmitt considers gender socialisation as a possible alternative to his hypothesis that gender differences in sociosexuality are innate. His discussion of the socialisation hypothesis, however, is rather concise. Upon reviewing three studies that find no evidence that men and women in Western societies hold explicit sexual double standards, he concludes that the socialisation hypothesis has been disproved. Based on this reasoning, he infers that we should not assume "that men and women will soon become equally promiscuous in both attitudes and behaviors, even when women are eventually treated as the social equals of men across all cultures" (Schmitt 2005, 272). In other words, Schmitt concludes from his brief encounter with social explanations that gender differences in sociosexuality are most likely "biological" and therefore unlikely to be affected by our efforts to build more gender equal societies.

In light of the discussion in this section, it is obvious that this chain of reasoning is highly problematic. Schmitt's decision to reject the socialisation hypothesis on the basis of three studies finding no evidence for explicit sexual double standards is not only based on a dubious evidential basis (how were those studies selected? What about the wealth of feminist literature on sexual double standards?) it also reflects an overly simplistic understanding of the complex and often subtle mechanisms of gender socialisation (Beal 1994; Grusec & Hastings 2007; Rudman & Glick 2010). Part of this oversimplification is the assumption that double standards for promiscuity in men and women do not exist merely because they could not be measured in people's explicit attitudes. As argued above, people can internalise such values subconsciously, without being able to report having them. As a result, the evidence brought forward by Schmitt is a long way from providing a conclusive case against the socialisation hypothesis.

In addition to this, Schmitt believes that the lack of evidence for explicit sexual double standards undermines both the socialisation hypothesis *and* the social structural ("structural powerlessness") hypotheses as plausible explanations for the observed gender differences (Schmitt 2005, 272). But this is not the case. Even if the cited studies would undermine the socialisation hypothesis as a plausible explanation of gender differences in sociosexuality, it may still be possible to explain the differences with social structural

features. Proponents of a social structural explanation for gender differences in sociosexuality could point to the fact that men and women are subject to very different social structural constraints when it comes to relative power in relationships, economic dependence on a long-term partner, or the risks of becoming victim of violence at the hands of a sexual partner. In light of these differences, they could argue that sex outside committed relationships is an objectively far more risky endeavour for women than for men. According to this argument, what discourages women from engaging in sociosexuality is not confined to the realm of cultural norms and ideas. It partly stems from real structural asymmetries which distribute the dangers associated with promiscuous behaviour unfavourably towards women.

5.3.3 Summary

In this section, I argued that it is important to distinguish two types of social explanations for gender differences: socialisation and social structural explanations. Socialisation explanations explain gender differences by pointing to differences in gender socialisation, that is, differences in social norms, values, and expectations that men and women learn about during their lifetime. While socialisation explanations always refer to learning processes, these learning processes can lead to different results, including conscious and subconscious internalisation as well as merely instrumental awareness of the relevant norms. Social structures are empirically intertwined with socialisation processes. Nevertheless, social structural explanations do not need to postulate learned differences in men's and women's psychology in order to explain gender differences in behaviour. Instead, social structural explanations can explain gender differences in behaviour by pointing to the different social constraints faced by men and women.

While evolutionary psychologists show some awareness of these distinctions, such as the difference between socialisation and social structural effects, they often do not fully acknowledge them in their arguments. This is crucial, because different types of social explanations come with different evidential requirements. I showed that the failure to recognise and address different types of social explanations for gender differences, has led some evolutionary psychologists to prematurely conclude that gender differences must be innate.

5.4 HOW EVOLUTIONARY PSYCHOLOGICAL AND SOCIAL EXPLANATIONS RELATE

The previous sections distinguished different types of social and evolutionary psychological explanations and explored to what extent the opposite parties recognise them. We saw that either side has only partial awareness of the explanations for gender differences offered by the other. Proponents of social explanations typically understand evolutionary psychologists as claiming that gender differences are innate or "hard-wired". They fail to acknowledge that some evolutionary psychologists instead claim that gender differences are purpose-specific adaptations, which may be compatible with proximate social explanations. At the same time, evolutionary psychologists demonstrate only limited awareness of the different types of social explanations, which sometimes leads them to premature conclusions. While this discussion already pointed to some ways in which social and evolutionary explanations may or may not be compatible, a systematic examination of how different combinations of social and evolutionary psychological explanations relate is still outstanding. In this section I explore how exactly each different type of evolutionary psychological explanation identified in Section 5.2 relates to each different type of social explanation explored in Section 5.3.

5.4.1 Explanatory relations

For that purpose, it is useful to first of all recall the different types of logical relationships that can obtain between two explanations. The following list is by no means exhaustive, but it outlines the options that are of main relevance to the debate.

5.4.1.1 Mutual exclusion

Two explanations are mutually exclusive if they make conflicting claims about the causal facts that bring about a phenomenon. Mutual exclusion implies that the two explanations cannot both be true (though they can, of course, both be false). The idea that social and evolutionary psychological explanations of gender differences are mutually exclusive has been popular amongst feminists as well as among evolutionary psychologists who, so I argued, implicitly understand evolutionary psychological explanations as claims about trigger innateness. Saying that two explanations are mutually exclusive is the same as saying that they are incompatible.

5.4.1.2 Compatibility

Compatibility is the opposite of mutual exclusion. To say that two explanations are compatible is simply to say that they are not mutually exclusive. In other words, two

compatible explanations do not make conflicting claims about causal facts and can therefore both be true. Two explanations can be compatible simply in virtue of being entirely irrelevant to each other, typically because they are not addressing the same phenomenon. For instance, the explanation that the moon is bright because it reflects the light of the sun is compatible with the explanation that homosexuality became decriminalized in the US because of the Stonewall riots, but this is simply because the two explanations are not at all relevant to each other.

This is not usually the case in the discussion on gender differences. Because the social and evolutionary explanations in questions tend to be explanations of the same type of phenomenon (for instance, observed gender differences in promiscuity) compatible explanations will generally be relevant to a joint explanatory purpose. Explanations which are compatible in this sense may differ with respect to explanatory or pragmatic concerns, that is, their underlying idea of a good explanation for the context at hand. Nevertheless, they can converge on the same causal-ontological story. One example of this form of compatibility are explanations which differ only with respect to causal selection. Such explanations implicitly recognise the same set of causal conditions for the occurrence of a certain phenomenon, yet differ in terms of which causal conditions they decide to foreground as explanatorily relevant in specific context. Depending on our interests, for instance, we may explain the fire in the barn by the presence of oxygen, or by the fact that the children were playing with matches. But both explanations converge on a single causal history, a history which involves both the presence of oxygen and the fact that children were playing with matches.

5.4.1.3 Complementarity (of proximate and ultimate explanations)

In the context of the discussion on gender differences, it is sometimes said that evolutionary psychological and social explanations are not mutually exclusive because they address different types of questions. This view, as in Section 5.1, presupposes that evolutionary psychological explanations are ultimate explanations which tell us why the extant population of humans consists of individuals who have a particular trait, and that social explanation are proximate explanations which describe the mechanisms by which individuals come to develop the trait during their lifetime. It is important to note that the claim can be interpreted in two different ways. The weaker interpretation merely states that evolutionary psychology's ultimate explanations are compatible with social scientist' proximate explanations, without making any suggestion as to how they relate. In general, however, proponents of this idea have a stronger claim in mind. They suggest that social

and evolutionary psychological explanations are *complementary* proximate and ultimate explanations of the same type of mechanism. For this to be true, the proximate mechanism that figures in a social explanation must have been subject to the evolutionary pressures described in evolutionary psychologists' ultimate explanations. The discussion on how social explanations relate to claims about purpose-specific adaptations will tell us more about this. But to begin with, I will first look at the relationship between social explanations and explanations about trigger innateness.

5.4.2 Social explanations and trigger innateness

5.4.2.1 Socialisation and trigger innateness

The relationship between socialisation explanations and trigger innateness explanations is seemingly straightforward. Since trigger innate traits are explicitly defined as traits which are not the product of learning, so it is commonly suggested in the debate, a trait cannot be trigger innate and the product of socialisation at the same time. Since the relevant contrast is with learned traits as such, we do not need to distinguish between conscious or subconscious internalisation and instrumentally learned traits. The common understanding suggests that socialisation and trigger innateness refer to two incompatible developmental processes – it's either one or the other, but not both. However, once we take a closer look, this description turns out to be a false dichotomy. After all, there is no reason why a particular gender difference (such as differences in mate preference) could not be the joint product of a trigger innate and a socialisation component. For instance, men's observed preference for young partners could be a trigger-innate tendency amplified by being socialised in a culture which objectifies women and values youthfulness.[3] In other words, socialisation and trigger innateness point to independent causal processes. But, despite a common misconception to the contrary, there is no reason to think that they are necessarily mutually exclusive.

5.4.2.2 Social structures and trigger innateness

A similar consideration applies to the relationship between social structural explanations and claims about trigger innateness. Recall that trigger innateness explanations of gender differences suggest the existence of different innate mechanism in men and women. These differences would result in different manifest behaviour even in identical social environments. Social structural explanations, by contrast, suggests that gender-differentiated behaviour is the result of differential structural constraints. These different

[3] Note that, in this case, it would of course be wrong to say that the *observed trait* is innate. Instead, we would have to say that the observed trait has an *innate component*.

social constraints, so the idea goes, will lead to gender-differentiated behaviour even if men and women do not at all differ psychologically – neither due to socialisation nor as a matter of trigger innate differences. We can see from this that social structural explanations and trigger innateness explanations refer to two processes for gender-differentiated behaviour which are causally independent – the one can occur without the other, and vice versa. However, this does not imply that social structural and trigger innate mechanisms cannot jointly be responsible for a particular behavioural outcome.

To see this, it helps to look in a bit more detail at the potential interactions between socialisation and social structural mechanism. Social structural mechanism, as suggested above, can work in tandem with socialisation. It is plausible to assume, for instance, that the structural reasons that keep women in the home are amplified by a gender role socialisation which portrays men as breadwinners and women as domestic and caring. How exactly can we conceptualise this "amplification"? I suggest a useful way of thinking about it is in terms of "tipping points". The tipping point, in this example, is the point at which structural constraints are so unequal as to affect differential behavioural responses from men and women. To illustrate this point, imagine we start off with a society of perfect gender equality where women are just as likely to pursue paid labour as domestic work. Assume that we gradually introduce a wage gap into this society. It seems likely that the more the wage gap disadvantages women, the more women will end up in domestic work rather than full-time paid employment. The same applies, mutatis mutandis, to men.

Socialisation interacts with these structural effects by changing individuals' tipping points between different labour arrangements. Gender-differentiated socialisation can influence men's and women's readiness to give up paid employment and take on child care. This, in turn, affects how big the wage gap has to be to motivate individual couples to adopt a certain division of labour. If women, for instance, are socialised into a caring role and men are not, couples might tend to adopt a traditional division of labour even in light of a narrow wage gap. If the socialisation of men and women emphasises care work equally, by contrast, it will probably take a wider wage gap to bring about the same pattern of division of labour.

Understanding this interaction between social structural and socialisation effects is important, because the relationship between trigger innate factors and social structural effects can be modelled in the same way. To see this, consider the suggestion that women, on average, have a stronger innate preference to care for children than men. This hypothetical innate difference can interact with structural effects just in the same way as

the learned difference described above. Hence, we can think of innate difference in caring preferences as influencing the tipping point of structural inequalities. In a society where women have a stronger innate preference for care work, the width of the wage gap required to make couples adopt a traditional division of labour would be narrower than in a society where men's and women's innate preferences are identical.

This discussion suggests that explanations which refer to psychological gender differences – be they innate or learned – are, in principle, perfectly compatible with social structural explanations. It is possible that social structural explanations and innateness explanations simply refer to different aspects of the causal history behind a specific gender difference.

5.4.2.3. Innateness, social explanations, and evidence

According to the previous discussion, any given gender difference could, in principle, be the joint product of any combination of socialisation, social structures, and trigger innate features. This suggests that social explanations and claims about trigger innateness are generally compatible. However, while the previous discussion suggests that both socialisation explanations and social structural explanations are in principle compatible with trigger innateness explanations on the level of *causal facts*, a tension becomes apparent once we consider how these explanations relate *evidentially*. Claims about trigger innateness, as argued in Section 5.2, need to be established by poverty of the stimulus arguments. Poverty of the stimulus arguments require demonstrating that the trait developed in an environment which is informationally impoverished relative to the output trait. This implies that the evidential route for establishing claims about trigger innateness is primarily an indirect or negative one: to show that a trait is innate, one needs to demonstrate that it appears in the absence of stimuli which are "informationally rich" with respect to the relevant trait. By contrast, if there is reason to believe that the environment is informationally rich enough to explain the trait without recourse to trigger innate components, claims about trigger innateness cannot get off the ground.

In the context of the debate on gender differences, this means proponents of trigger innateness explanations would have to argue that the social environment is informationally impoverished relative to the observed gender-differentiated behaviour or psychological feature. In other words, they would have to demonstrate that the nature of the observed gender differences cannot be explained by reference to socialisation or social structural effects alone. The problem for proponents of trigger innateness explanations, then, is that both socialisation and social structural explanations of gender differences

describe environments as informationally enriched in the relevant respects. In socialisation explanations, the relevant informational content is present in cultural messages about social norms, values, and expectations and can be consciously or subconsciously internalised or just instrumentally learned. In social structural explanations, the informational content is imprinted into structural constraints which provide men and women with different options at different costs. This means that, as long as we have reason to believe that any of these social explanations are true for a specific gender difference, we have no reason to think that there could be a trigger innate component involved.

Evolutionary psychologists make some efforts to show that the environment is informationally impoverished in the relevant ways. In Section 5.3, for instance, we saw Schmitt arguing that gender differences in sociosexuality must be innate because there are no sexual double standards in modern Western nations. Yet, due to Schmitt's failure to distinguish socialisation from social structural mechanisms, this attempt fails. In other words, if evolutionary psychologists are to make a thorough case for trigger innate gender differences, they would have to engage with different types of social explanations of gender differences to a much greater extent than they presently do. In a surprising argumentative twist, evolutionary psychologists sometime try to turn this reasoning on its head. Hence, Buss argues that

> [the] structural powerlessness hypothesis […] and subsequent social structural variants are fundamentally indefensible, because *their core premise of male and female identity* of underlying psychology was always theoretically problematic and is now known to be empirically false
>
> (Buss 2005, 279, my emphasis)

However, such claims about the empirically established falsehood of the innate psychological identity of men and women are not generally backed up with any evidence other than evolutionary psychological conjecture. Buss' chain of reasoning, for instance, concludes that evolution must have created innate, gender-specific adaptations on the psychological level simply because it has created them on the anatomical level as well. The upshot of Buss' argument, then, is that social structural (and, presumably, socialisation) explanations of gender differences have no plausibility unless they demonstrate that there are no innate psychological gender differences. In other words, he suggests that the requirement of a poverty of the stimulus argument for trigger innateness is mirrored in a requirement for a "poverty of innate psychology" argument for social explanations.

This would be true if proponents of social explanations, by using the same reasoning technique as evolutionary psychologists, simply took the very existence of gender differences as evidence that there is a social explanation for it. Social scientists might use this technique as a heuristic device for discovering gender-differentiated aspect of socialisation or social structures that previously went unnoticed. But in general, social explanations of gender differences draw on a much richer source of evidence which often directly relate to the specific sort of mechanisms they describe. Consider the following, very brief sampling of such empirical resources: Studies show that adults evaluate and interact with babies differently depending on what they assume the baby's gender to be (Culp et al. 1983; Meyer & Sobieszek 1972; Sobieszek 1978). Systematic evaluation of the presentation of men and women in media, literature, and advertising supplies ubiquitous evidence for gender-differentiated cultural norms and expectations (Davis 2003; Smith 1994; Lauzen et al. 2008; Kay & Furnham 2013). Evidence for differential social structural constraints faced by men and women is provided, for instance, in the form of statistical data on the wage gap, professional segregation, or women's higher susceptibility to becoming a single parent, to working in precarious conditions, or to becoming subject of violence at the hands of an intimate partner (Equality and Human Rights Commission 2011; Office for National Statistics 2015; Department of Health 2005).

In other words, unlike claims about innate psychological traits, claims about socialisation or social structures are based on empirical evidence for the underlying mechanisms. Such evidence undermines evolutionary psychologists' claims about innateness, because it undermines poverty of the stimulus arguments. The result is an evidential asymmetry between trigger innateness explanations on the one hand and social explanations on the other. While proponents of social explanations can base their explanations directly on empirical evidence, proponents of trigger innateness – unable to produce any direct evidence for their own hypotheses – rely crucially on the absence of evidence for social explanations to give any credit to the idea that the mechanisms they propose exist. But the same is not true the other way around. Even if we had evidence for the existence of innate psychological gender differences, this would not by itself undermine social explanations of gender differences as long as there is direct evidence for the existence of the proposed social mechanisms.

5.4.3 Social explanations and purpose-specific adaptation

Having clarified the relationship between social explanations and claims about trigger innateness, it is time to consider how social explanations relate to the arguably less

controversial interpretation of evolutionary psychologists' claims: the idea that gender differences are purpose-specific adaptations. On this interpretation, to say that a trait is the product of domain-specific psychological mechanisms is to say that it is an adaptation or, more precisely, that there has been selection for the trait in question because it increased our ancestors' fitness. The relationship between social explanations and purpose-specific adaptation claims is bound to differ in one important respect from the relationship between social explanations and trigger innateness explanations. Trigger innateness, socialisation, and social structures all point to proximate mechanisms for the production of human behaviour. Hence, the relationship between these different explanations is a question of how different proximate mechanisms relate to each other. Claims about purpose-specific adaptations, by contrast, are claims about ultimate causation. As explained previously, they concern the selective history of a trait. They tell us whether a trait is widespread in the human population today for a particular reason – because it made a specific contribution to our ancestors' fitness.

As a result, there is a specific range of possible answers to the questions how social explanations and purpose-specific adaptation claims relate. In principle, there seem to be two ways in which social explanations and claims about purpose-specific adaptation could relate. The social explanation could describe the proximate mechanism *for* realising the adaptive trait. In this case, the social explanation would pick out factors which contribute to the same mechanism which produced the trait in our ancestors and which have been subject to natural selection. In this case, the social explanation and the evolutionary psychological explanation would be complementary, because they provide proximate and ultimate explanations of the same phenomenon.

Alternatively, social explanations may describe a proximate mechanism for the trait in question which could not plausibly have been subject to natural selection. In this case, the social explanation and the explanation in terms of purpose-specific adaptation are not complementary, but they may still be compatible. The reason for this is that there might be other mechanisms which contribute towards the development of the trait and which have been subject to natural selection. In the debate at hand, a prime example of this scenario would be a gender difference which is the joint product of non-adaptive social mechanisms and an adaptive innate component. As long as there exists a component which has been favoured by natural selection for contributing to the gender difference in question, it would be correct to say that the gender difference is a purpose-specific adaptation. Since the social mechanism, in this scenario, also causally contributes to the

gender difference, we could say that the difference is socially caused *and* a purpose-specific adaptation, even though the two explanations are not in any way related.

In other words, social explanations and claims about purpose-specific adaptations can be complementary, and they can be compatible without being complementary. Could they also be mutually exclusive? Evolutionary psychologists, as we have seen above, sometimes suggest that social and evolutionary psychological explanations are compatible simply in virtue of the fact that they address different levels of causation (proximate versus ultimate). If evolutionary psychological explanations refer to claims of purpose-specific adaptations (rather than trigger innateness in disguise) this is indeed the case. If social explanations of gender difference are compatible with the idea that the difference is in part due to innate psychological differences between men and women, it is also compatible with the claim that the difference is a purpose-specific adaptation. The reason for this is that the innate psychological difference could be the product of natural selection. As a result, social explanations and claims about purpose-specific adaptations are not generally mutually exclusive, but either compatible or complementary. Before closing this chapter, we need to consider under what circumstances different types of social explanations may be complementary to, rather than merely compatible with, claims about purpose-specific adaptations.

5.4.4 Socialisation and purpose-specific adaptation

When asking whether socialisation explanations could be complementary to claims about purpose-specific adaptation, we are essentially asking whether natural selection can operate on psychological traits which are learned rather than trigger innate. The reason for this, as argued above, is that proximate and ultimate explanations are complementary if the mechanism or factor identified in the proximate explanation is the same mechanism or factor which is identified as the product of natural selection in the ultimate explanation. Since socialisation explanations explain gender differences in terms of social learning processes, they can only be complementary to ultimate explanations if natural selection acts on transmission by learning.

Incidentally, this latter question has been much discussed in philosophy of biology, with the general answer being "yes, but only under specific circumstances". In practice, there seem to be a number of human and non-human traits whose development, as well as transmission across generations, crucially relies on social learning. Prominent examples include language, upright gait, tool use and, arguably, imitation (Heyes 2011; 2018; Ray & Heyes 2011; McNeil et al. 1984; Sterelny 2005; Laland 2004). Moreover, all of these traits

are likely to make an important contribution to an individual's survival and reproduction. It is therefore plausible to assume that these traits are adaptations which are transmitted via learning. Eva Jablonka and Marion Lamb have explored this idea in detail on a more theoretical level (Jablonka & Lamb 2005). They argue that it is possible for natural selection to act on traits which are transmitted by social learning rather than genes. This is not obvious, because transmission by social learning differs from transmission by, say, genes in one crucial respect. Unlike genes, social learning not merely allows vertical (parents to offspring) transmission, but also for horizontal (within-generation) transmission. Horizontal transmission tends to undermine natural selection, because it allows traits to spread in the population irrespectively of their fitness benefit.

This does not mean that horizontal transmission is automatically irrelevant to the discussion at hand. Some would argue that horizontal social transmission can give rise to an evolutionary process in its own right called *cultural evolution* (Richerson & Boyd 2005; Kronfeldner 2011; Godfrey-Smith 2012; Lewens 2015). Cultural evolution describes a process in which socially transmitted traits spread in the population, not (or not primarily) depending on whether they increase the fitness of those who carry them, but depending on whether they possess "cultural fitness". However, it is plausible to assume that the concept of cultural selection is irrelevant to the discussion at hand. The reason for this is that cultural fitness describes a trait's success in being transmitted across a population independently of whether or not it conveys a fitness advantage onto the individual who possesses the trait. Evolutionary psychologists, however, are concerned with whether or not a trait has spread in the population because it conveyed a fitness advantage on individuals who possess it. This means that, on evolutionary psychologists' understanding of adaptations, socially transmitted traits that have been subject to cultural rather than natural selection to not qualify as purpose-specific adaptions.

All this suggests that, in order for a socialisation explanation to be complementary to an ultimate explanation, the socialisation explanation must refer primarily to vertical (parent to offspring) rather than horizontal transmission by learning. But this is not generally the case for socialisation explanations of gender differences. Socialisation explanations of gender differences typically suggest that the norms, values, and expectations which underlie gender-differentiated socialisation permeate large parts of society. They are not only, and possibly not even primarily, communicated by parents, but also by peers, the media, teachers, and so forth. In other words, in the case of gender socialisation, horizontal transmission of the relevant norms, values, and expectations is

ubiquitous and might well overpower vertical transmission. This suggests that socialisation explanations do not usually describe mechanisms which could have been subject to natural selection. As a result, we should not expect socialisation explanations to be complementary to the ultimate explanations put forward by evolutionary psychologists.

5.4.5 Social structures and purpose-specific adaptation

Finally, consider the question whether social structural explanations and claims about purpose-specific adaptations can be complementary. As before with socialisation explanations, another way of phrasing this question is to ask whether social structural explanations refer to the same proximate mechanisms that evolutionary psychologists identify as purpose-specific adaptations. We can see that this idea runs into difficulties even more easily than when applied to socialisation explanations. Natural selection, as argued above, depends on their being heritable (that is, vertically transmittable) differences between individuals. Social structures, quite obviously, are not heritable differences between individuals. Instead, they are environmental factors that, by definition, affect individuals in the same social positions in the same way. While socialisation mechanisms could, at least in theory, be subject to natural selection acting on vertical transmission by learning, this is not the case for social structural effects.

In principle, this leaves open the possibility that social structures could be subject to a different type of selection. Some philosophers and evolutionary biologists recognise the possibility of *group selection* (Griesemer & Wade 1988, Sober & Wilson 1998). Group selection is a form of natural selection acting on groups not individuals, and favouring the evolution of traits which are adaptive for the group rather than the individual. It might be possible to conceptualise social structural mechanisms as (socially) heritable differences between human groups, for instance between different cultures. In this case, social structural mechanisms might qualify for something like group selection acting on socially transmitted variants. To defend this option, we would have to discuss in more detail the controversy on group selection and explore how it may apply to social structures (Lloyd 2001; Okasha 2006). However, these efforts would not help us to clarify the debate on gender differences. The reason for this is that evolutionary psychological explanations of gender differences are concerned only with individual-level adaptations, that is, traits which have been selected for increasing the fitness of individuals rather than groups. The issue of group-level selection is irrelevant to their claims about purpose-specific

adaptations. As a result, there is no reason to expect social structural explanations to be complementary to ultimate explanations.

5.6 CONCLUSION

This chapter illustrated and shed some light on the common phenomenon of hybrid kinds standing at the crossroads of social and natural scientific explanations. It did so by examining the hitherto poorly understood relationship between social and evolutionary psychological explanations of gender differences – a crucial and long overdue task in its own right. I argued that we need to distinguish several types of causal claims on each side. On the side of evolutionary psychological explanations, it is possible to identify three types of claims: claims about domain-specificity, claims about trigger innateness, and claims about purpose-specific adaptations. I argued that evolutionary psychology's central concept of domain-specific psychological mechanisms, and the way it is used in discussing social explanations, resonates only with the idea of trigger innateness and the idea of purpose-specific adaptions. As a result, only trigger innateness and purpose-specific adaptation were considered in the further discussion.

Turning to social explanations, I distinguished socialisation explanations from social structural explanations, and identified three different types of socialisation explanations. I pointed out that, although socialisation and social structural effects are intricately linked and partly mutually constitutive, distinguishing them is essential for two reasons: socialisation and social structural explanations rely on different types of evidence, and they may relate to evolutionary psychological explanations in different ways.

Finally, I explored how the different types of social and evolutionary psychological claims identified in previous sections relate to each other. The answers varied and some were quite unexpected. The discussion revealed that social explanations of any type are in principle compatible both with trigger innateness explanations and with the claim that gender differences are purpose-specific adaptations. At the same time, we should not generally expect social explanations to provide proximate explanations that are complementary to these ultimate explanations. The finding that stands out most, however, is that no combination of social explanations and evolutionary psychological explanations makes mutually exclusive claims about causal facts. On a more general level, this suggests that natural and social scientific explanations for hybrid kind phenomena may tend to be compatible on the level of causal facts simply in virtue of the fact that they describe causal factors which may act in concert.

Given the heated debate on explanations for gender differences, this finding was somewhat puzzling. However, we discovered that this conflict may be rooted in an *evidential* rather than causal tension. The tension, so I argued, arises between trigger innateness explanations and basically any type of social explanation. It stems from the fact that the empirical evidence used to support social explanations for a specific gender difference will generally undermine trigger innateness explanations for the same gender difference because it precludes poverty of the stimulus arguments. This gives rise to an evidential asymmetry between social and evolutionary psychological explanations which has interesting implications for the debate on gender differences. At least on a theoretical level, proponents of social explanations of gender differences have little to fear from evolutionary psychology. Since proponents of social explanations for gender differences can usually draw on data that directly supports claims about the existence of the relevant mechanisms, there is no scientific need for them to challenge the existence of innate gender differences. These findings, however, leave open the question what to make of evolutionary psychological explanations for gender differences if one's stakes in the debate are political rather than scientific. The next and final chapter of my thesis explores this issue on a more general level, by asking whether biological explanations of gender differences are relevant to feminist politics.

6

CAN FEMINISTS IGNORE BIOLOGY?

The previous chapter explored what happens when hybrid kinds, in virtue of their structure of base kind and associated social status, simultaneously fall within the explanatory realm of both natural and social sciences. But hybrid kinds not only tend to stand at the crossroads of natural and social scientific explanations. As argued in Chapter 3, the question whether or not an individual occupies a specific classification-associated status can have a crucial impact on their lives. As a result, hybrid kinds also have a tendency to stand in the crossfire of science and politics. This last chapter explores how the political-pragmatic concerns that can be associated with a hybrid kind's status relate to explanatory concerns regarding its base kind. Since the previous chapter suggested that there are few cases where this tension is as pronounced as in the discussion on gender differences and gender inequality, I will continue to focus on this case study.

More often than not, the battle lines in relation to the debate on gender differences run between social scientific and biological approaches of a certain type. The biological explanations in question are explanations which suggest that men and women differ "naturally", that is, independently of specific social and cultural arrangements. They typically involve claims supporting existing stereotypes about the role and behaviour of men and women, for instance that men are more competitive and women more caring. These biological explanations are usually juxtaposed in opposition with social explanations, that is, explanations for gender differences that make recourse to differences in social learning inputs or social structural arrangements for men and women (as illustrated in Chapter 5). In other words, when feminists critically refer to "biological" explanations of gender differences, this should not be understood as an outright rejection of any explanation employing biological vocabulary. Instead, such claims are best understood as addressing those explanations which propose an innate basis for observed psychological and behavioural gender differences, in a way that is seen to confirm existing stereotypes about men and women.[1] Despite the risk of sounding overgeneralising, I will stick to the expression of "biological" explanations of gender differences for the sake of simplicity and to keep in line with the common terminology in the debate.

[1] For a specification of the concept of innateness, and an argument that many evolutionary psychological explanations of gender differences are best interpreted as proximate claims about innate differences, see Chapter 5.

With this qualification in mind, we can explore in more detail the ways in which feminists tend to sympathise with social scientific explanations of gender differences while being wary of biological ones. Some feminists – usually those who hold a background in biology or neuroscience themselves – critically scrutinise biological explanations of gender differences quite extensively (Fausto-Sterling 1992, 2000; Fine 2010, 2017). They try to undermine these explanations by identifying methodological flaws and sometimes propose alternative frameworks and methodologies. This strategy is illustrated, to some extent, in Chapter 5. There, I argued that evolutionary psychologists' arguments that gender differences are trigger innate are often undermined by their insufficient engagement with social scientific explanations of the same phenomena. Most feminist theorists and activists, however, do not have an in-depth understanding or carry out detailed scrutiny of the science behind these claims themselves. Instead, some of them only refer to existing criticisms pragmatically, to reject biological arguments for gender differences where they come their way (Oakley 1972). Others go further than that and do not seem to engage with biological explanations of gender differences at all (MacKinnon 1989; Okin 1989; Haslanger 2012; Asta 2012; Jenkins 2016).

In other words, there is a significant number of feminist theorists and activists who take a keen interest in social scientific explanations of gender differences but deliberately ignore biological ones (Lucal 2010). This happens to the dismay of many proponents of biological explanations, who lament the "biophobia" amongst feminist and other left-leaning social theorists and activists (Pinker 2003; Campbell 2002; Barkow 2006). Some go so far as to liken large parts of today's feminist movement to climate change deniers (Tierney 2016). Many of them urge feminists to engage with the relevant biological research, usually under the promise that it will benefit their efforts at furthering gender equality, but generally to little avail.

This chapter investigates the adequacy of this prevalent feminist disinterest. Are proponents of biological explanations right in their demand for being heard, or do feminists have perfectly good reasons for ignoring them? I start this investigation by considering Maria Kronfeldner's "right to ignore", which centres on defending the autonomy of cultural anthropology from biological anthropology. Section 6.2 identifies parallels and differences between Kronfeldner's discussion of cultural anthropology on the one hand and the discussion on gender equality on the other, and concludes that Kronfeldner's argument cannot be applied to the feminist case. Section 6.3 examines two further strategies for justifying feminists' ignoring of biological explanations. Finding

problems with both, it leaves the task of identifying an alternative justification, which will be developed in Section 6.4.

6.1 KRONFELDNER'S RIGHT TO IGNORE

6.1.1 Alfred Kroeber and the right to ignore evolutionary anthropology

Luckily, our inquiry into whether feminists are justified in ignoring biological explanations of gender differences does not have to start from scratch. A case with some striking parallels has been discussed in the context of cultural and evolutionary/biological anthropology (Kronfeldner 2009, 2017; Meloni 2016a, 2016b). Unlike our debate, which concerns explanations for differences between men and women, the discussion in anthropology has traditionally been fought over how to explain differences between human racial groups. Evolutionary anthropologists used to argue that differences in human cultural achievements reflect innate cognitive differences between human racial groups. Cultural anthropologists, by contrast, explained these differences in terms of further social and cultural differences.

But in order for this to happen, cultural anthropology first had to emancipate itself from evolutionary anthropology in the early nineteen hundreds. According to Maria Kronfeldner, one of the founding fathers of cultural anthropology, Alfred Kroeber, played a crucial role in this process. Kroeber, Kronfeldner suggests, successfully argued that cultural anthropologists have a *right to ignore* evolutionary anthropology's claims about "natural" racial differences in their research (Kronfeldner 2009). By defending this right, Kroeber adopted what Kronfeldner calls a "separationist epistemic stance". The separationist epistemic stance (or simply "separationist stance") is a research heuristic that defends the right to ignore certain phenomena or causal factors in explanations typical for a specific discipline (Kronfeldner 2017). Kronfeldner argues that the separationist stance is an undervalued research heuristic which is usually overshadowed by a "synthesis" bias amongst both scientists and philosophers. The synthesis bias consists in the – according to Kronfelder unjustified – believe that synthesis and integration of scientific theories and perspectives are the best recipe for scientific progress. Kronfeldner wants to show that this bias is unfounded, and that separation and integration can be equally epistemically valuable research strategies.

Kronfeldner makes two important qualifications to the right to ignore. Firstly, she emphasises that the right to ignore goes both ways. While her case study specifically shows how cultural anthropologists may be justified in ignoring evolutionary anthropologists'

claims about natural differences, evolutionary psychologists may be just as justified in ignoring cultural anthropologists' claims about cultural differences (Kronfeldner 2017, 214). In addition, she points out that the right to ignore does not consist in "*willfully accepting known inconsistencies*, but in *not checking for consistency*" with biological explanations (Kronfeldner 2017, 214, original emphasis). It is important to keep these points in mind, because we will come back to them later.

6.1.2 Epistemic and ontological arguments for the right to ignore

Kronfeldner presents two arguments to support the right to ignore, an epistemic argument and an ontological one (Kronfeldner 2017; 2018). The epistemic argument aims to challenge the synthesis bias by demonstrating that a separationist stance can be just as epistemically fruitful as an "integrationist stance". According to Kronfeldner, proponents of an integrationist stance argue that we need to integrate different theories and explanations because doing so is *epistemically fruitful* – it produces new insights, theories, and even disciplines. Kronfeldner uses the case of Kroeber to demonstrate that a separationist stance can be just as conducive to reaping these epistemic benefits. She suggests that Kroeber's separationist stance produced several major epistemic benefits.

The first is the establishment and keeping alive of cultural anthropology as an autonomous discipline with its own methods. The second is the combating of scientific racism, which was based on the idea of Lamarckian inheritance (more on this below). In addition to that, Kronfeldner credits Kroeber with developing the precursor of a new scientific field –dual or multiple inheritance theory. Dual/multiple inheritance theory is a flourishing approach that examines the interaction of different systems of inheritance, most notably genetic and cultural inheritance (Jablonka & Lamb 2005; Richerson & Boyd 2005). Some people even suggest that the insights from multiple inheritance theory, together with other advances in evolutionary and developmental biology, suggest a novel conceptual framework for thinking about evolution (Laland et al. 2015). According to Kronfeldner, although Kroeber's strategy is too radical today, as we recognise important interactions between nature and culture in human evolution, his realisation that biological and cultural inheritance can be considered as two distinct systems of inheritance was a precondition for these theoretical developments.

While the epistemic argument is intended as a general defence of the separationist stance as a fruitful research heuristic, the ontological argument pertains specifically to the subject matter of anthropology and the question whether the nature-culture that informs Kroeber's right to ignore is ontologically plausible. Kroeber, as argued above, defends the

right to ignore evolutionary or biological claims about innate psychology when explaining human culture. This argument suggests that there are aspects of human psychology and behaviour which can be explained in terms of culture alone, without making reference to facts about human biology or genetics. Some people, however, would argue that this claim is untenable on the ontological level. Generally speaking, critics object that the influences of nature and culture are too entangled to be usefully singled out, and that persistent attempts at doing so are misguided and only hamper scientific progress. They object that any human trait is the product of complex interactions between genetic, epigenetic, and environmental factors. As a result, it simply does not make sense to argue about whether a certain human trait is largely "due to the genes" or "due to the environment". Instead, all human traits are the joint product of genetic and environmental factors, and their individual contributions cannot be apportioned.

This line of argument is usually made with regard to the nature-nurture debate, which concerns the relative influence of genetically inherited and environmental factors on human development (Fox-Keller 2010; Tabery 2014; Taylor 2014). Since "environmental factors" include cultural ones, the overall idea applies just as much to the nature-culture divide.[2] Natural and cultural factors, so the idea goes, are entangled in such a way that we cannot meaningfully identify their relative contributions. This would seem to undermine the argument for a mutual right to ignore between cultural and evolutionary anthropology. If every single human trait or behaviour is the joint product of biology and culture, we would expect the biological factors investigated by evolutionary anthropologists to be relevant to cultural anthropology and vice versa.

Against this view, Kronfeldner insists that we can identify an ontological distinction, on the basis of which it is possible to defend the right to ignore: the distinction between biological inheritance on the one hand and cultural inheritance on the other. According to Kronfeldner, we can meaningfully distinguish these forms of inheritance because they involve two largely independent channels of transmission which are characterised by distinct internal dynamics (Kronfeldner 2018). While Kronfeldner develops this argument in more detail, and in a way that accommodates the recent discoveries on epigenetic inheritance, she points out that the general idea can already be found in Kroeber's account of culture as "superorganic" (Kroeber 1917, cited in Kronfeldner 2017). According to Kronfeldner, Kroeber used evolutionary biologists' rejection of Lamarckian inheritance (the inheritance of acquired characteristics), in order to counter the then common belief

[2] For a discussion that focusses specifically on the nature-culture divide, see Oyama (2000).

that "cultural change is evidence for and is causally linked to biological evolution" (Kronfeldner 2009, 117). Without inheritance of acquired characteristics, Kroeber argued, culture can still influence culture. For instance, the cultural ideas that exist in a society at a given time may influence people's values and behaviour, which likely influences how the society develops culturally. But contrary to what evolutionary anthropologists at the time used to believe, a society's culture cannot influence the innate and heritable cognitive capacities of those living in it. As a result, Kroeber concluded, culture is "superorganic" in the sense that cultural change is independent of biological evolution, in particular change in innate cognitive abilities.

6.1.3 How do the epistemic and the ontological argument relate?

In Kronfeldner's view, then, Kroeber realised that culture and nature can be conceptually distinguished as separate systems of change and inheritance and defended his right for cultural anthropologists to ignore nature on the basis of this observation. However, it is not quite clear from Kronfeldner's discussion how exactly Kroeber derived this right to ignore from the distinction between cultural and biological inheritance. In particular, Kronfeldner does not clarify to what extent the epistemic success of ignoring innate cognitive capacities as a fruitful research heuristic depends on the existence of an underlying ontological distinction between biological and cultural inheritance. This points us to the more general problem that Kronfeldner does not spell out how exactly the ontological argument and the epistemic argument relate in establishing a right to ignore. Intentionally or not, Kronfeldner's discussion thus leaves open the possibility that the epistemic argument alone can establish a right to ignore – that there might be cases where a separatist stance historically proves to be epistemically fruitful even if we cannot find ontological reasons for why this is the case. Such a "success proves right" position might be defensible, but it comes with certain caveats that make it have limited use in the context of the discussion on gender differences.

The most obvious problem with the "success proves right" argument is that it tells us what we are epistemically justified in ignoring only after we have already ignored it. In the context of science, this usually means that we have to build methodologies and theories on the assumption that we are justified in ignoring a certain aspect of our target phenomenon before having any idea as to whether doing so is a good idea. For anyone who stands at the outset of a new research project on gender differences and wonders whether it would be epistemically fruitful to ignore certain factors associated with the target phenomenon, the "success proves right" argument is of little help. It provides no

guidance on how to determine heuristically which factors could be ignored to our epistemic benefit, and therefore leaves us with a pure trial-and error strategy. Of course, it is perfectly possible that scientific progress is made on the basis of this pure trial-and error strategy. But the approach is obviously costly, and we would probably prefer an alternative strategy that gives us more guidance, if one exists.

The problems with relying on the "success proves right" argument may go even deeper than that. Assume a group of scientists practice ignoring specific factors (think of feminist social scientists ignoring the possibility of innate cognitive differences in explaining of gender differences) and is epistemically successful on the basis of this research strategy. Without identifying any underlying ontological characteristics that suggest a separationist stance is likely to be successful in this specific context, we have no way of telling whether the epistemic success was merely coincidental. After all, we are not given any reason as to why a research strategy taking into account innate cognitive differences might not have led to the same or even superior epistemic results. This brief discussion does not demonstrate that the separationist stance cannot be justified on the basis of a "success proves right" argument alone. However, it suggests that we will fail to understand when and why a separationist stance is likely to be successful unless we take a keen interest in the ontological side of the story. In other words, it is time to take a closer look at Kronfeldner's ontological argument and its relationship to the epistemic success of Kroeber's separationist stance.

Kronfeldner's ontological argument, as explained above, states that Kroeber's right to ignore potential innate cognitive differences between racial groups is justified because biological and cultural inheritance can be distinguished as two largely independent channels of inheritance. How exactly does this argument work? According to Kronfeldner, the missing link in the argument stated above is that scientific racism at the time was based on a Lamarckian view of evolution, that is, the idea that acquired characteristics are heritable. While innate physical differences between human groups could be readily observed, evidence for innate cognitive differences had to be conjured up by sophisticated argumentation. The Lamarckian view of evolution offered just that. On this view, one could argue that innate racial differences in cognition exist because human racial groups live in radically different environments. In response to these different environments, so the idea, different racial groups develop different physical as well as cognitive characteristics. These characteristics are then inherited, accumulate over time

and explain racial differences in cultural achievements, in particular "higher" (Western) and opposed to "primitive" (non-Western) cultures.

Kronfeldner's suggests that, without the Lamarckianism interaction between biological and cultural change – and methods for searching for racial differences directly in the genome not yet available – there is no reason to assume innate racial differences in the first place. Observed cultural differences between racial groups no longer support scientific racism, because they can be neither understood as evolutionary causes nor as developmental symptoms of innate cognitive differences. Instead, Kronfeldner suggests, scientists would have to presuppose what Kroeber called the "psychic unity of mankind." This assumption states that although there may be individual innate differences in cognitive abilities, we should expect them to average out on the level of racial or cultural groups.

As Kronfeldner observes, the assumption of psychic unity plays a crucial role in Kroeber's argument that cultural anthropologists can ignore claims about innate cognitive differences. If innate cognitive differences average out at the level of racial and cultural groups, they are clearly unsuitable for explaining cultural differences between these groups. In Kronfeldner's words, "nature becomes a disciplinary primitive [for cultural anthropology]: one can safely assume it and then ignore it since it does not make a difference for the historical change in culture that the cultural anthropologist aims to explain" (Kronfeldner 2017, 215). This means that, under the assumption of psychic unity, cultural anthropologists can safely ignore innate biological differences because they are irrelevant to what they want to explain. In other words, Kroeber did not deny that there is a shared human nature, or individual differences in natural abilities, but he denied that either of these are relevant for explaining what he wanted to explain: differences in shared culture. Instead, Kroeber came to the conclusion that "culture explains culture" – cultural differences across human groups can only be explained by reference to pre-existing cultural differences (Kronfeldner 2009, 104-5).

There is a further aspect of Kronfeldner's paper that requires mentioning. In some passages, Kronfeldner suggest that a crucial aspect of Kroeber's strategy was to reconstitute culture (understood as a system of inheritance and change in its own right) as an explanandum (Kronfeldner 2009, 123). According to Kronfeldner, this idea can be contrasted with earlier accounts which understood culture either as a different kind of explanandum, such as patterns of behaviour and symbols (Tylor 1871, cited in Kronfeldner 2009) or as an explanans of human behaviour rather than an explanandum

in its own right (Boas 1911, cited in Kronfeldner 2009). In particular, Kronfeldner seems to suggest that Kroeber's strategy of ignoring nature was only feasible because his explanandum was human culture, not human behaviour.

This difference, however, is overstated. On the assumption of psychic unity, the lack of a reason for assuming group-level natural differences affects explanations of group-level behavioural differences just as much as group-level cultural differences. Neither cultural nor behavioural group differences can be explained by reference to natural differences unless there is reason to assume that natural between-group differences exist. This implies that the heavy lifting in Kroeber's argument for a right to ignore natural differences is done not by reconstituting culture, understood as a system of change and inheritance, as the explanandum. Instead, the heavy lifting is done by the assumption – based on insights from evolutionary biology – that there are no relevant innate cognitive difference between human racial groups. The logic of this argument applies to group level behavioural differences just as much as to cultural differences.

Taken together, these considerations provide a clear answer to the question as to how the epistemic and the ontological argument for the right to ignore relate. Ignoring factor F (such as innate cognitive differences) is likely to be epistemically fruitful if it is possible to define the explanandum in such a way that there is good reason to assume that factor F makes no difference to what we want to explain. Whether it is possible to define the explanandum in this way, however, depends crucially on the ontological structure of the target domain – the reason innate cognitive differences can be ignored in explaining group cultural differences is because innate cognitive differences are unlikely to accumulate at the level of cultural groups.

This suggests that, strictly speaking, Kronfeldner's right to ignore does not have that much to do with ignoring after all. While Kronfeldner admits that her argument does not establish a right to ignore inconsistencies that we are aware off, it does not establish a right to not check for inconsistencies we might be unaware of either. The reason for this is that such a right would be redundant in Kroeber's example. We already know that there aren't any such inconsistencies, because we have defined the explanandum in such a way that – given the assumption of psychic unity – inconsistencies with claims about innate cognitive differences are impossible. In other words, the strategy defended by Kronfeldner does not consist in ignoring inconsistencies, but in reasoning them out of existence.

6.2 FROM RACE TO GENDER – PARALLELS AND DIFFERENCES

6.2.1 The psychic unity of "humankind"?

In order to find out whether Kronfeldner's right to ignore can be transferred to the discussion on gender differences, we need to examine to what extent the two cases are comparable. To begin with, we can immediately note a crucial disanalogy between Kroeber, who ignored the question of innate cognitive differences among racial groups, and feminists who ignore the question of innate cognitive differences between men and women. Kroeber's right to ignore is based on his observation that the state of the art biology of his time provided no reason to believe in innate cognitive differences between racial groups. In more recent times, this idea has been corroborated again and again by scientific and technological advances. Despite scientists' persistent attempts to provide a scientific foundation for entrenched folk racial classifications, geneticists and philosophers have insisted that none of their suggestions captures salient ontological structures (Lewontin 1972; Lewontin et al. 1984; Cavalli-Sforza et al. 1994; Gannett 2004, 2005; Bolnick 2008; Kaplan & Winther 2013). The only plausible exception to this are the social structures of often gaping inequalities that have formed around folk racial classification, and the imprint that these structures tend to leave on the biological, for instance in the form of health disparities (Kaplan 2010). In other words, the assumption that there are no innate cognitive differences between racial groups continues to prevail not only within cultural anthropology. It has become widespread consensus in virtually all academic disciplines dealing with human difference.

Once we turn to the question of innate cognitive differences between men and women, the situation is very different – certainly in Kroeber's time but still very much today. Kroeber's crucial assumption of the psychic unity of "mankind" was primarily that – the assumption that there are no innate cognitive differences between men of different cultures. Implicitly, this assumption may suggest that there are no innate cognitive differences between women of different cultures either. But it says nothing on the matter of innate cognitive differences between men and women in general. It would be difficult to interpret this silence on the matter as anything other than agreement with the common belief of Kroeber's time: that there are innate differences between men and women, across all cultural and racial groups, and that they make women suitable to traditional subordinate roles and men to traditional dominant ones.

It was not until Ruth Benedict and Margaret Mead that the common assumption of innate cognitive differences between men and women began to be questioned within

cultural anthropology (Mead 1928; 1932; 1949; Benedict 1934). However, unlike their colleague Kroeber, who could invoke evolutionary theory to support his argument for ignoring claims about innate racial differences, Benedict and Mead could not draw on support from the biological sciences. To the contrary, the idea that there are innate cognitive differences between men and women and that these are relevant to explaining behavioural differences was common sense at the time and has remained a respectable idea until today. Presumably, this is partly due to the fact that evidence which has become available to challenge the idea of innate cognitive differences between human racial groups simply is not suitable to challenge the same idea with respect to differences between men and women. Although we know today that, much like racial categories, *male* and *female* cannot be understood as non-overlapping categories with necessary and sufficient properties for membership, crucial differences between sex and racial categories remain.

The most important one, I suggest, concerns to what extent these categories have proven to be epistemically useful. There is good reason to believe that sex categories qualify as homeostatic property cluster (HPC) kinds. HPC kinds are scientific categories which are constituted by a cluster of properties that are reliably coninstantiated due to underlying homeostatic mechanisms (Boyd 1989). It is widely assumed that HPC kinds are scientifically useful because their reliably associated properties facilitate inductive inferences. I argued in Chapter 4 that this claim has to be qualified – the reason such kinds facilitate inductive inferences is because we typically have a decent grasp of the underlying mechanisms.

On either understanding of HPC kinds, there are good reasons to believe that sex categories fit the bill. While it is important to recognise the existence of intersex people – individuals whose combination of anatomical features (such as chromosomes, gonads and genitalia) does not conform to the mixes that we associate with either *male* or *female* – it would be wrong to conclude that this undermines the use of *male* and *female* as scientific categories in the human sciences. For one thing, the sets of properties that we associate with *male* and *female* are still very reliably associated. Combining numeric estimates for the frequency of different sex anatomy variations, the total percentage of intersex people is estimated to be 1.7% (Fausto-Sterling 2000a).[3] What's more, scientists understand the physiological mechanisms that support these patterns of association and variation to a

[3] Some people have objected that Fausto-Sterling's estimate is too inclusive and quantify the percentage of intersex people at 0.018% instead (Sax 2002).

remarkable extent (see Bashamboo & McElreavey 2016 for a review article). As a result, sex categories provide a useful base for inductive inferences in several scientific and practical contexts.

The same can hardly be said for racial categories. As argued above, although scientists persistently try to identify an empirical basis that supports a similar type of argument for racial categories, such claims remain highly controversial. Anatomical properties, so the objection goes, do not cluster any more within traditional racial groupings than within many other possible groupings of the human population. As a result, racial categories simply do not cut nature at one of its joints – not even a fuzzy cluster joint – and therefore do not make useful scientific categories.

These observations suggest that there is a crucial difference between the case of sex categories and racial categories when it comes to establishing a right to ignore. In the case of race, science has again and again undermined any reason for thinking that innate properties cluster within racial groups. The same is not true for sex categories. Since the assumption of a lack of innate cognitive differences at the group level played a crucial role in Kroeber's argument for a right to ignore, we should not expect that argument to carry over smoothly to the case of gender differences. As a result, Kronfeldner's account of a right to ignore may be valuable for critical race theorists, but it will be of little use to feminists.

6.2.2 From nature-culture to sex-gender

Against this background, it may come as a surprise that many feminist scholars have adopted a strategy that is, in some ways, remarkably similar to Kroeber's. The nature-culture distinction defended by Kroeber is, to some extent, mirrored in the sex-gender distinction commonly employed by feminists. Consider how the sex-gender distinction has been established. Feminists appropriated the concept of gender from sexologist John Money in the late 1970s. Money had introduced the term in the 1950s to distinguish physical characteristics of men and women from psychological and behavioural ones (Money 1955, 1957). But even before that, feminists interested in the social dimension of sexual categories used notions like "sex role" and "sexual identity" to demarcate their subject matter from sex, understood as a biological category. Since then, the distinction of sex role/sexual identity/ gender as opposed to physical sex has been used to identify a field-defining phenomenon for women's and gender studies (Germon 2009).

This process is strikingly similar to the establishment and demarcation of cultural anthropology on the basis of the nature-culture distinction. But more than that, scholars

in women's and gender studies tend to ignore biological explanations of gender differences, to the extent that terms like "biology" do not even feature in some gender studies dictionaries (Griffin 2017). If scholars of women's and gender studies engage with biological explanations of gender differences at all, it is often by way of explicitly contrasting their own approach to the biological study of gender (Martin 2008).

The question, then, is this: how can feminists working in the areas of women's and gender studies pursue the same strategy of ignoring innate cognitive differences as cultural anthropologists, without being able to draw on comparable scientific support for the assumption of a "psychic unity" between men and women? Since the conditions that enabled Kroeber to establish a right to ignore innate cognitive differences between racial groups are simply not given in the case of gender differences, it seems that we need to look for justification elsewhere.

Note that there is additional reason to look further afield. Kronfeldner argues that the right to ignore applies both ways. This would suggest that feminists can ignore biological explanations of gender differences just as much as biologists can ignore feminists' social explanation of gender differences. But this is at odds with feminists' dual strategy of criticising biological explanations methodologically while ignoring them in their own approach. This approach suggests that feminists rely on a one-sided right to ignore: while feminists seem to consider themselves justified in ignoring biological explanations of gender differences in theory and practice, biologists' failure to consider social and cultural explanations of gender differences is seen as both epistemically and politically dubious. In other words, what would be needed is a justification for ignoring biological explanations that does not depend on scientific evidence for the "psychic unity" between men and women, and which makes sense of feminists' conviction that their right to ignore is not a mutual one. In the following, I will consider two justifications which are commonly brought forward to this effect in the feminist literature.

6.3 FLAWED SCIENCE AND IDEOLOGY – HOW FEMINISTS JUSTIFY IGNORING BIOLOGY

Generally speaking, when feminists defend the decision not to engage with biological explanations of gender differences, they typically invoke either of the following two objections:

(i) Biological explanations of gender differences are based on flawed science.

(ii) Biological explanations of gender differences are just ideology.

The first objection is usually made by reference to one of the many existing critical discussions of biological explanations of gender differences (see, for instance, Fausto-Sterling 1992, 2000; Fine 2010, 2017). Authors of these discussions claim to have identified a number of methodological shortcomings and flawed reasoning in extant biological explanations of gender differences. In doing so, they argue that these explanations fall short of the standards that scientists are expected to set to themselves. This is why the "flawed science" objection provides a powerful tool for challenging proponents of biological explanations of gender differences on their own terms. Feminists who bring forward "flawed science" objections for ignoring biological explanations argue that they can ignore these explanations because everyone else should, too. The explanations in question, so they argue, are scientifically unfounded, and therefore do not have special epistemic status.

The second objection refers to the concept of ideology and is a bit more difficult to pin down than the "flawed science" objection. Philosophers have analysed the concept of ideology in much detail (see, for instance, Geuss 1981). To get an adequate understanding of the "just ideology" objection, however, it will suffice to understand the concept of ideology in a very general manner. Broadly speaking, "just ideology" objections do not (or not primarily) approach biological explanations of gender differences as claims of special epistemic status. This is because "just ideology" objections are typically made in a context where the focus is not so much on whether or not these claims are true, but on how they reflect and reinforce extant power structures in a society. This approach goes back to Michel Foucault and has been advanced in the context of feminism and gender studies by Judith Butler (Foucault 1978, Butler 1990). Feminists who make "just ideology" type objections tend to be more interested in how biological explanations reflect and reinforce the social circumstances from which they emerge (Contratto 2002, Cassidy 2007, Ahmed 2008, Lucal 2010). They understand biological explanations of gender differences first and foremost as another form of discourse which reflects and reinforces the values and distribution of power in a society.

In the context of this feminist ideology critique, questions about the truth of biological explanations can in principle be ignored because they are irrelevant to what the approach aims to investigate: the social and cultural factors with lead to the acceptance of biological explanations, as well as the social and cultural effects which result from the acceptance of

biological explanations. These processes, so the "just ideology" approach assumes, operate quite independently of whether the explanations are factually correct or not.

Both the "flawed science" and the "just ideology" objection stem from approaches that provide important insights into discussions of gender differences. "Flawed science" objections are based on an approach that provides detailed scientific scrutiny of biological explanations of gender differences. By improving the methodological soundness of biological explanations and preventing us from making political decisions on the basis of dubious scientific claims, it may yield scientific as well as political benefits. "Just ideology" objections operate within a framework that can provide important insights into the nature of the debate – such as the conditions under which scientific claims about the causes of gender differences become a recognised tool for negotiating power in a society – and may point to ways in which this discourse can be destabilised and resisted.

Although both approaches make highly valuable contributions to the debate, each has crucial limitations when it comes to justifying a wholesale right to ignore biological explanations of gender differences. The "flawed science" objection relies on specific points of criticism targeted at specific biological studies of gender differences. It thus lacks generalisability – the criticism levelled at one type of biological explanation does not automatically apply to the next, let alone to biological explanations of gender differences in general. This means the "flawed science" objection applies only to biological explanations which have been shown to be based on a flawed methodology or false assumptions, plus those explanations which are similar to these in relevant respects. It does not provide feminists with a justification to ignore biological explanations once and for all. To the contrary, it may demand detailed scientific engagement with every single scientific study which claims to have identified a biological basis for gender differences in behaviour or psychology. This is unhelpful as a general feminist strategy. If feminists had to comprehensively scrutinise each and every study that claims to have found biological determinants of gender differences for methodological soundness and plausibility of assumptions, they would have little time to do anything else.

The "just ideology" objection avoids this problem. Unlike the "flawed science" objection, it does not depend on the methodological and empirical details of individual scientific studies. Instead, the "just ideology" approach can in principle be applied, quite sweepingly, to any scientific claim about gender differences which is found to express or reinforce social inequality between men and women. But there are important drawbacks. The "just ideology" objection operates within a framework that can challenge claims

about gender differences only at a level that is external to the scientific discussion itself. It therefore does not allow us to challenge scientists on their own terms, and, most importantly, puts aside the question as to whether their explanations are true. To many proponents of biological explanations of gender differences, this will mean feminists simply sidestep the core question. They will object that, even if biological explanations of gender differences are symptomatic of women's oppression and used to justify it, feminists cannot refuse to engage with them as truth-apt claims. Proponents of this view typically motivate their position by saying that whether or not biological explanations are true will make an important difference to feminists' politics.

Proponents of the "just ideology" approach could respond that the truth of biological explanations of gender differences *is*, to a great extent, irrelevant to feminist concerns. The reason for this, they could argue, is that the widespread acceptance of such explanations will be harmful to women no matter if the explanations in questions are true or not.[4] In the extant social climate, so the idea, these explanations will predictably and inevitably be used to reinforce stereotypes about men and women, to disqualify women for positions of power, and to undermine efforts of addressing the social determinants of problems like sexual violence and the gender wage gap.

Proponents of biological explanations, however, will likely reply that the feminist response it too one-sided and pessimistic. They will point out that there might be important benefits to accepting and publicising biological explanations of gender differences that we have reason to believe are true. Most notably, they will suggest, it might help us to identify important opportunities for intervention – if we discover that gender-differentiated behaviour is socially flexible in certain respects – as well as limitations – if we discover that it isn't. This could allow feminists to target their political interventions more carefully and effectively. In other words, the problem with using the "just ideology" approach to justify ignoring biological explanations wholesale is the unclear net benefit. Proponents of biological explanations object that the benefits of engaging with biological explanations "properly" may well outweigh the costs. Although feminists have good reasons to assume that the promulgation of these explanations will be harmful to women in some respects, these harms could be overridden by the potential benefits that may come from the insights which biological explanations provide.

[4] For a more general argument that explanations of human behaviour can have harmful effects irrespectively of their truth value, see Mallon (2016, Chpt. 4).

In sum, neither of the two rationales commonly used by feminists is quite sufficient to legitimise a wholesale right to ignore biological explanations of gender differences. The "flawed science" approach only justifies the piecemeal ignoring of biological explanations that have been shown to be scientifically lacking. The "just ideology" approach in principle licenses a more comprehensive right to ignore biological explanations, on the ground that such explanations may tend to reinforce gender inequality. But it is vulnerable to the objection that feminists risk prematurely dismissing biological explanations which provide crucial insights into combating gender inequality. In the following, I will identify rationales for ignoring biological explanations that avoid both of these pitfalls. They explain why feminists can ignore scientific explanations of gender differences more generally, without making them vulnerable to the objection that they forego scientific knowledge which is essential to their cause.

6.4 A RIGHT TO IGNORE FOR FEMINISTS?

In this section, I develop two distinct justifications for ignoring biological explanations of gender differences. The first justification is based on feminists reconstituting *gender* as their explanandum. This strategy is facilitated by the hybrid nature of the kinds *men* and *women*, but has important limitations that will be described below. The second justification is based on pragmatic considerations about the relevance of biological explanation to a wide range of feminist purposes, and is not subject to the same limitations. I will discuss both in the following, starting with the strategy of making gender the explanandum.

6.4.1 Explaining gender and explaining gender differences

As mentioned in several previous chapters, many human classifications can be understood as hybrid kinds. They consist not only of what I call a "base kind", constituted by the properties that define the category, but also of an associated "status kind", constituted by the social position that individuals acquire qua being recognised and treated as members of the specific category. A fruitful way to understand the concept of gender is in terms of this hybrid kind model. On this view, the feminist distinction between sex and gender attempts to conceptualise the hybrid nature of the categories *men* and *women*, with "sex" denoting the base kind and "gender" the associated status kind. Gender, in other words, refers to the social statuses – as opposed to the bodily features – that are associated with the human classifications *men* and *women*.

This observation has crucial implications on the question whether feminists can ignore biology. Status kinds are characterised by a position in a network of social relations. They

are constituted by how one is regarded and treated by other people, and possibly by oneself. This means that the properties which characterise a status kind are exclusively social, relational properties. By redefining their target phenomenon in terms of the status kind gender, feminists gain considerable autonomy from biological explanations of sex differences. The reason for this is that the connection between base kind and status kind is peculiar and fragile. As argued in Chapter 1, the social position that individuals come to occupy in virtue of being classified has little to do with the properties that these individuals have prior to being classified. Members of the base kind come to occupy the social position that characterises the status kind only if they are recognised as members of the base kind. At the same time, individuals who are wrongly believed to be members of the base kind nevertheless come to occupy the associated social position, if only enough people who stand in relevant relations to them share the false believe. In other words, the connection between base and status kind is highly conventional, because it entirely depends on whether or not we (consciously or subconsciously) recognise individuals as members of a certain kind.

This makes gender, in principle, independent of sex. According to the hybrid kind model, the fact that individuals classified as men or women occupy specific social status positions is not the result of their biological makeup. Biological characteristics are merely markers on the basis of which individuals are singled out for being treated and thought about in certain ways. The only link between their prior properties and their social position goes through us.

This suggests that feminists can make use of a strategy similar to Kroeber's after all. As cultural anthropologist did with culture, feminists can define their explanandum (gender) in such a way that makes biological explanations irrelevant for explaining it. In fact, it seems that the concept of gender can establish an even more far-reaching right to ignore biology than Kroeber's concept of culture, because the success of the argument does not depend on biological facts about innate psychological differences. This is because culture and gender are independent of biological explanations for different reasons. Kroeber's argument for the independence of culture from biology essentially depends on a biological assumption about the psychic unity of human racial groups. The argument for the independence of gender from biology, by contrast, does not depend on biological but on social facts – the fact that gender, as a status kind, is attached to specific biological properties only by convention.

On the downside, and precisely because psychic unity cannot be assumed, the right to ignore biology by distinguishing sex from gender has a serious limitation. It is limited to explanations of gender, and does not extend to explanations of behavioural and psychological differences between men and women. This is a big price to pay in the debate at hand. While there may be gender studies scholars who are interested in gender on a purely theoretical level, for instance by exploring and comparing the different statuses women occupy in different cultures, the feminist concern with gender usually goes beyond that. Feminists ultimately want to use the insights from investigating gender to change society, which means changing the way people think and act. In order to further gender equality, one could argue, it would seem that feminists need a robust understanding of why men and women tend think and act differently in patriarchal societies. If this is true, behavioural and psychological differences between men and women are a relevant explanandum for feminist purposes and cannot simply be defined out of sight.

6.4.2 Do biological differences matter?

The remaining question, then, is whether feminists may be justified in ignoring biological explanations of behavioural and psychological gender differences. To answer this question, I suggest that we start by turning the tables and ask: what reasons do feminists have for paying attention to these explanations? One of the responses that immediately springs to mind is that paying attention to biological explanations prevents feminists from making claims that are factually wrong. The response, prominent for instance among proponents of evolutionary psychology, is often motivated by the epistemological assumption that "hard" sciences like biology provide a more reliable guide to the causal structure of the world than the "airy" social sciences and humanities, together with the ontological assumption that, since biological and feminist explanations of gender differences differ, at least one of them must be wrong.

Both assumptions have been discussed in some detail in Chapter 5. Against the ontological assumption, I argued that evolutionary explanations (often understood as claims about innate gender differences) and feminist social explanations need not necessarily conflict. Instead, both may correctly identify factors that contribute to gender-differentiated behaviour. With regard to the epistemological assumption, I casted doubt on the idea that the epistemic strengths of feminist social explanations versus biological explanations can be judged in such a general manner. I argued that claims about innate differences, for instance, are usually established by poverty-of-the-stimulus arguments. To establish such arguments, proponents of biological explanations would need to

demonstrate that the social causes identified by feminists are causally insufficient, but their studies regularly fall short of this requirement. As a result, the scientific superiority of biological explanations cannot simply be presumed in a wholesale manner. Instead, different types of explanations are supported by different methodologies and different forms of empirical evidence, and their relative empirical adequacy and methodological soundness have to be established on a case-by-case basis.

The idea that paying attention to biology prevents feminists from making false claims is closely related to another common justification for why feminists should consider biological explanations of gender differences. According to this justification, biological explanations are of immediate practical concern to feminists concerns because they help identify which aspects of social life can be changed and how easily. This idea has already been widely discussed, so I will not dwell on it for long (see, for instance, Lewens 2003; Dupre 2001; Buller 2005). The widespread consensus states that human behaviour and psychology are highly complex and plastic, and the only reliable way to find out whether a certain intervention will be effective is to try it out or consult experts on the influence of relevant environmental factors. This suggests that knowledge of biological causes of gender differences (such as innate cognitive differences) will likely tell us little or nothing about whether and how they can be changed. In addition, feminists are rarely interested in intervening on biological causes, but tend to focus on social ones (more on this below). In sum, there seems to be little reason why feminists should consult biological explanations of gender differences in order to identify relevant levers for change.

Seeing as knowledge of biological explanations is far less important for understanding and altering gender differences than proponents of biological explanations have made it out to be, what other reason could there be for feminists to pay attention to such explanations? I think there is a third widespread motivating assumption, which, unlike the previous two, is rarely spelled out. According to this view, knowledge of biological causes of gender differences is relevant to feminists not because it tells us *how* to change unjust or oppressive social arrangements, but in order to determine *which* arrangements are unjust or oppressive in the first place. The argument underlying this claim can be spelled out as follows: Inequalities which are due to innate differences are not a matter of justice because they are not the result of social causes. Hence, if feminists want to claim that observed inequalities between men and women are unjust and need addressing, they first need to establish that these inequalities are the product of social factors rather than innate differences. On this view, knowledge about the biological determinants of gender

differences is essential for feminist purposes not primarily because of its epistemic or practical relevance. It is essential because the existence of such biological determinants has crucial normative implications and could undermine feminists' justice claims.

The view at hand is most clearly expressed in Janet Radcliffe-Richards' discussion on sex equality (Radcliffe-Richards 2014, 45; see also Radcliffe-Richards 1998). Radcliffe-Richards criticises what she calls "only X% arguments", and which have the following form:

> Justice demands sexual equality.
>
> But only x% of CEOs/ senior academics/ government leaders…. are women; Women have only x% of male earnings/ leisure time…; Men do only x% of housework/ child care….
>
> Therefore there is still unjust inequality between the sexes.
>
> (Radcliffe-Richards 2014, 45)

According to Radcliffe-Richards, conclusions of arguments of this type do not follow, at least not if we evaluate them on the basis of what she considers one of the most fundamental and uncontroversial accounts of equality (Radcliffe-Richards 2014). This idea, Radcliff-Richards argues, is today known as "ground-level impartiality" or "equal considerations of interests" but ultimately goes back to John Stuart Mill's concept of "perfect equality". The principle of perfect equality prohibits the *arbitrary* disadvantaging of one group merely for the purpose of advantaging another. This principle can be applied in the context of whatever other rules and principles organise life in a society. In Radcliffe-Richards' words, it is about "removing [...] balls and chains" that are arbitrarily attached to certain players in a game without questioning the principles of the game itself. As a result, Radcliffe-Richards points out, the principle is very powerful and very limited at the same time. It is very powerful because of its generality. According to Radcliffe-Richards, the principle is "now effectively beyond controversy" and therefore constitutes an "essential aspect of sexual justice" (Radcliffe-Richards 2014). Simultaneously, it is very limited because it does not allow us to question any further principles that guide the distribution of wealth and power within a society.

Radcliffe-Richards suggests that adopting Mill's principle has crucial implications for "only x%" arguments: it implies that observed inequalities between men and women count as unjust only if they are the result of arbitrary differential treatment rather than due to biology. Consider the following example. Within a liberal capitalist society that allows for significant inequalities between individuals based on "merit", the principle dictates that it would not be permissible to arbitrarily hinder women from entering

capitalist positions of power. It would not be permissible, for instance, to systematically socialise girls in ways that make them less qualified than boys for a wide range of positions of power, unless such gender-differentiated socialisation could be justified on independent grounds. If it turned out, however, that women are *innately* less suitable for these positions (for instance because they are innately more caring and less competitive), the resulting underrepresentation of women in positions of power would be perfectly in accord with Mill's principle of perfect equality. In other words, the strength of feminist complaints about only x% of women occupying positions of power depends on whether the underrepresentation of women is caused by social rather than by biological differences.

But Radcliffe-Richards criticism of "only X%" arguments goes further than that. She insists that the burden of proof about what causes women's underrepresentation lies on the side of those who claim that there are no innate gender differences. Since men and women differ in "systematic and striking ways", Radcliffe-Richards argues, "no scientist would decide that it was reasonable to presume they must be alike, on average, in unknown ways unless there was positive evidence to the contrary" (Radcliffe-Richards 2014, 55). She concludes that the case for discrimination based on "only x%" observations is only as strong as the evidence that men and women "are intrinsically alike in all relevant respects" (Radcliffe-Richards 2014, 54-55).

This is of course controversial. We have encountered in previous chapters a fairly extensive literature suggesting that matters regarding the burden of proof and the state of empirical evidence for innate gender differences are much more complicated than Radcliffe-Richards admits. But the argumentative strategy she invokes is widespread and has proven quite resistant to this criticism. For that reasons, it is worth looking at the discussion from the other end and ask: what really follows for feminist politics if we accept Radcliffe-Richards's argument? *Pace* Radcliffe-Richards, I suggest that her argument, if correct, would still provide an excellent justification for feminists to ignore claims about biological differences. The reason, put simply, is as follows. If innate differences are not a matter of gender equality – as Radcliffe-Richards insists they aren't – there is no reason why feminists should consider them.

To see this, consider alternative explanations as to why knowledge of such differences is relevant to feminists. One suggestion would be to say that knowledge of biological gender differences is pragmatically important for targeting feminist interventions – if feminists want to rectify biologically caused gender inequality, they need to understand the

biological pathways by which these inequalities come to be. But this is implausible. Innate, biological pathways for gender-differentiated behaviour or psychology are generally not something that feminists tend to mobilise against or seek to manipulate in order to further gender equality.

Consider a notorious example. If we really had reason to assume that innate hormonal differences make women on average less apt at maths than men, it would be quite unusual for feminists to argue that hormone levels need to be adjusted to further women's prospects in the labour market. A more plausible feminist response would be to object that the prestige of jobs involving mathematical ability reflects a male-biased standard that devalues professions traditionally carried out by women, such as care work. But this would involve feminists campaigning for a better remuneration and change of public perception of traditionally female occupations – not for chemical "correction" of hormonal differences. In other words, since intervention on biological causes is not something that feminists of any persuasion tend to advocate, it would be wrong to say that they need knowledge of biological differences for that purpose.

Opponents of a feminist right to ignore biology could reply that the above suggestion is disingenuous. Knowledge of biological differences, they could argue, plays a rather different role in arguments about gender equality. The reason feminists need to understand biological determinants of gender differences is not because they may want to manipulate biological pathways, but because, epistemically, it is needed to determine whether the resulting inequalities are unjust in the first place. This view relies on the Millean principle of perfect equality and combines it with the implicit assumption that any evidence for observed gender differences being due to innate differences is at the same time evidence against the suggestion that such differences are due to social factors. If we accept both assumptions, evidence for innate gender differences would undermine the idea that the observed inequalities between men and women are unjust because only inequalities that stem from arbitrary differential treatment constitute injustices.

The argument suggests that, in order to establish that a certain gender inequality is unjust, feminists need to make sure that it is not caused by innate differences. It assumes furthermore that the best way of doing so is for feminists to engage with biological claims about innate differences. This reasoning is seen as so self-evident by proponents of biological explanations of gender differences that it is are hardly ever spelled out. But it is far from obvious that it is true. Developmental pathways are numerous and complex. As argued in Chapter 5, different factors are not generally mutually exclusive, but can act

additively or interact in more complex ways. Accordingly, the idea that differential social treatment contributes to gender-differentiated behaviour is not made any less plausible by evidence for innate gender differences.[5] In this context, trying to establish the efficacy of a certain causal factors by excluding the efficacy of all others would be a very poor research heuristic. A far more common and fruitful approach for establishing causal claims is by investigating causes individually to find evidence for their efficacy. But this is exactly what happens in large parts of feminist social scientific research. Researchers in this field have found a plethora of evidence for the differential expectations, attitudes, judgements, and behaviour that people demonstrate towards women and men (see, for instance, Brewer 2001; Fine 2010; Ridgeway 2011; Weisgram & Dinella 2018).

This discussion suggests that, although we cannot simply assume that any observed gender difference is due to social factors, a different sort of reasoning is perfectly valid: to assume, on the basis of considerable evidence for the pervasive differential treatment of men and women, that observed psychological and behavioural gender differences and inequalities are at least partly attributable to social factors. Accordingly, in order to establish that gender differences are at least partially caused by social factors, feminists have little reason to consider claims about biological differences. All they need to do is produce evidence for the existence of the relevant social factors. Add to these observations Radcliffe-Richards' claim that arbitrary differential treatment constitutes one of the most fundamental and uncontested forms of injustice and we suddenly have a substantial case for feminists to ignore claims about biological difference. As a result, even within the narrow framework set out the Millean principle of perfect equality, feminists can identify discriminating social arrangements, and make their abolishment a matter of justice, without giving much consideration to biology. As long as feminists have robust evidence for the existence of social determinants of gender inequality, there is no need to consult a biologist before identifying the inequality as unjust.

6.4.3 Moving beyond Mill

The argument above illustrates that a feminist right to ignore biology can be defended even when making significant concessions to the opponent. As explained above, Mill's principle requires equal treatment within existing social structures but leaves the structures themselves unquestioned. But the greater – and, arguably, more insightful – part of

[5] See Chapter 5 for an argument that the same is not true the other way around, i.e. the belief in innate gender differences is methodologically undermined by evidence for gender-differentiated socialisation or gender-differentiated social structural positioning.

feminist thought and activism is concerned with the critique of those very structures. As a result, few feminists worth their salt would limit themselves to the Millean principle. To put things into perspective, it is therefore important to briefly consider the range of feminist concerns that falls outside this minimal liberal justice framework and explore how they are affected by questions of natural difference. I will start with fairly obvious cases and then move on to the more tricky ones.

To begin with, we can note that numerous feminist justice claims depend on the acknowledgement, rather than the denial, of natural differences between men and women. In key discussions such as those concerning birth control, abortion, or maternity leave, all parties agree that biological gender differences exist and that they are relevant to the discussion. This does not in any way diminish their status as matters of justice. To the contrary, feminist arguments in these areas are based on the fact that women's bodies are generally naturally different from men's in certain respects. Feminists argue that social institutions have traditionally been designed with male bodies in mind and that justice demands addressing and accommodating these natural differences. Accordingly, feminists argue that female hygiene products should be tax-exempt, that women should have access to affordable birth control and abortion, and that pregnancy and breastfeeding should be adequately accommodated in the workplace and public life more generally. In other words, there is a wide range of feminist debates in which acknowledgement of natural differences does not undermine justice claims but rather underpins them. The reason acknowledgment of natural difference is taken to support justice claims in these arguments is that most feminists find the Millean framework of perfect equality unjustifiably restrictive. They argue that a feminism which blindly accepts existing "rules of the game" is not a feminism worth having and that critique of social structures should be at the heart of the feminist project (see, for instance, Foster 2016).

In addition to cases where natural differences support rather than undermine justice claims, there are several core feminist issues for which the question of natural differences has clearly very little relevance at all. Consider the discussion on fair pay for domestic and care work. Unless one were to propose that women have an innate desire to do unpaid labour, the question of natural difference is entirely peripheral to arguments for fair compensation for care and domestic work. Feminist campaigning on this matter typically relies on the idea that such work makes a crucial contribution to capitalist production and as such deserve adequate compensation. Questions such as whether women have more natural inclination or innate talent to do care work are quite immaterial to this argument.

What matters is that care and domestic work are necessary and quantifiable contributions to the economic output of a society.

Similar considerations apply with regard to feminist criticism of the objectification of women in the media. The claim that women are widely objectified in movies, TV and advertising, and that this objectification is harmful on several levels, does not require recourse to questions of natural differences. It is a directly observable fact, and its harmful consequences have been widely documented by psychologists (Fredrickson & Roberts 1997; Moradi & Huang 2008; Carr & Szymanski 2011). Both examples are central matters of feminist discourse and campaigning. In both cases, the argument that women are treated differently from men in a way that disadvantages them and facilitates further oppression can be made by considering social facts alone. Questions of innate cognitive differences simply do not arise, or if they do, are not relevant to determining the injustice in question.

So far, we have seen that there are feminist issues where natural differences clearly do matter, but support rather than undermine justice claims, as well as feminist issues where natural differences clearly do not matter at all. In addition to these two fairly straightforward types, there is a third type of more complex cases. In the feminist discussion known as the "porn wars", for instance, the extent to which questions of innate differences are relevant may not be immediately obvious (Bronstein 2011).[6] Some feminists argue that the depiction of women in porn is harmful because it encourages sexual violence against women (MacKinnon 1987). The reason for this, they suggest, is that porn normalises and reinforces sexual aggression and the sexualisation of dominance. This claim can be juxtaposed with evolutionary psychologists' suggestion that innate disposition may play a role in explaining male-against-female sexual violence (Thornhill & Palmer 2000).

Feminists who argue that porn causes sexual violence often point to psychological studies which suggest that watching porn increases relevant attitudes, and criminal statistics which suggest that sexual offences tend to be inspired by porn (Scully 1990, Malamuth et al. 2000). Empirical evidence for these claims is contested (see Ferguson & Hartley 2009), but we can put the question of empirical adequacy aside for now. The purpose of this section is not to settle empirical debates, but to explore, on a conceptual and pragmatic level, the respective roles of claims about social causation ("porn causes

[6] Note that this section employs a narrow definition of pornography as erotica that sexualises the subordination of women. While pornography constitutes the bulk of erotica, this definition leaves open the possibility of erotica without subordination.

sexual violence against women") and claims about biological causation ("innate male disposition causes violence against women") with regard to feminist political demands. Let's assume then, for the sake of the argument, that empirical evidence clearly shows that consumption of porn encourages sexually violent behaviour in men. Many feminists would admit that this observation does not generally allows us to make inferences about whether or not men (more so than women) have an innate tendency towards sexual aggression. The reason psychological studies and crime statistics cannot disprove evolutionary psychologists' claim, many feminists would argue, is because sexual violence and the eroticisation of domination are ample in the societies under study.[7] As a result, we lack a comparison group. Since studies and statistics inevitable work with subjects that have been brought up in societies that sexualises dominance, we cannot determine – or, some would argue, even imagine – what sexuality would look like without these influences.

Defendants of the porn industry could argue that this methodological concession backfires against the feminist critique. They could insist that determining the extent of "innate sexual violence" is elementary to the discussion at hand, because we cannot put the blame on porn without some way of quantifying how much harm done to women is actually due to pornographic imagery and narratives. Since we lack the control group of a society free of images and stories that normalise and romanticize sexual violence, it is impossible to know whether innate sexual psychological features would produce similar levels of sexual violence against women in the absence of porn. This information, they could argue, is vital because agents like the porn industry can only be held accountable for whatever harm they are actually responsible for, i.e. the difference between these two scenarios.[8]

There are several ways feminists could respond to this. Firstly, they could object that asking "how much" sexual violence women would suffer in a society rid of porn means engaging in a fruitless and misleading form of social atomism, the attempt to see social problems as distinct matters that can be addressed individually. Pornography, they would object, is not an isolated issue. It is both symptom and reinforcing factor of women's oppression in its many ugly shades. Accordingly, sexual violence should not be understood as produced by pornography alone, but as the result of a comprehensive

[7] To be precise, feminists could argue that this observation undermines innateness explanations *methodologically*, i.e. it undermines the idea that evolutionary psychologists have sound evidence for their claims (see Chapter 5). They could not, however, infer the *ontological* claim that male innate tendencies towards sexual violence do not exist.

[8] This discussion does not take into account women who are harmed by porn directly, that is, the atrocities of coerced production. Arguments against porn made on the basis of direct harm are unaffected by the discussion above.

system of oppression that involves economic, political and cultural aspects. Since all these aspects are interrelated, trying to imagine a society that lacks porn but leaves all other aspects of gender oppression intact makes little sense both from an epistemic and from a political perspective. One could potentially compare the status quo to a utopian society in which women's oppression has been eradicated altogether. But this strategy, if it is possible at all, would obviously fail to answer the "how much harm due to pornography" question. The reason for this is that the utopic control group is one in which a whole range of factors other than porn have been manipulated. It is not a suitable control group for the question at hand.

The upshot of this discussion, then, is as follows. With phenomena such as porn, that are part of complex systems of oppression, the whole attempt of quantifying how much harm is done or who owes what to whom may be misguided. Because systems of oppression are so pervasive and complex, we may not even be able to imagine what an appropriate control situation would look like, let alone have any reason to assume that such a thought experiment will provide an effective tool for alleviating the relevant social problems. There is, in other words, much reason to believe that comparing the existing state to a hypothetical ideal situation is neither necessary nor helpful for understanding and ameliorating existing social problems. For that reason, many feminists and other social theorists are refusing to engage in this so-called "ideal theory" approach (see Mills 2005). Still, the question remains how feminists can criticise a phenomenon like porn without being able to specify what exactly the removal of porn would achieve, or how the problem can be rectified without overthrowing all of women's oppression at once. But this question loses its urgency once we recognise that it is not necessarily the goal of critical feminist analysis to tell us how we can improve specific situations or who owes what to whom. Instead, the primary – and in the foreseeable time only realistic – aim of such critique may be to raise awareness of the role that aspects of everyday life play in systems of oppression.[9]

In addition, in the case of porn at least, it may still be plausible to make the following case. If there is robust evidence that porn increases violence towards women, it is irrelevant, from a moral point of view, whether this happens by reinforcing existing psychological tendencies or by creating new ones from scratch. In the light of this evidence, postponing action until the harm done by porn has been meticulously quantified reflects a concern for the producers and consumers of porn that can only be interpreted

[9] See Finlayson (2016) for a detailed argument to this effect.

as a complementary disregard for those who suffer from it. Why not think that knowledge *that* porn harms would be enough to consider ways of regulating pornography, holding producers and distributors to account, or altering public perceptions of sexuality? In either case, there is little reason to think that feminists could further their agenda more effectively by consulting biological explanations. While social psychological research on whether and how pornography harms women is immediately relevant to this project, research into innate psychological inclinations towards sexual violence is not.

6.4.4 Summary

The preceding discussion suggests that the prospects of a feminist right to ignore biological explanations are much more promising than the discussion in previous sections may have made us believe. Depending on the explanadum in question, different justifications are available. If the explanandum is gender, understood as a status kind, biological explanations can be disregarded for the simple reason that they are irrelevant by definition of the explanandum. If the explanandum is the behaviour and psychology of men and women (and potential differences thereof), the situation is slightly more complicated. I identified two candidate reasons for why the information provided by biological explanations could be relevant to feminists. It may be relevant to understanding whether and how gender differences can be changed, or it may be relevant because it tells us whether observed inequalities are really the objectionable injustices feminists make them out to be.

I noted that the first candidate reason has already been extensively discussed in the extant literature and ultimately been rejected. Claims about innate differences, so a widespread consensus states, are not generally relevant to understanding if and how behaviour and psychology can be changed. I then showed that the second candidate reason, which has not received a comparable amount of critical attention, is equally problematic. Although Radcliffe-Richards is right to point out that "only X% arguments" cannot generally be used to establish claims about discrimination, this does by no means imply that feminists need to consider biological explanations each time they want to identify a social arrangement as unjust. To the contrary the question of innate differences in entirely immaterial to several key issues in feminism, and of highly debatable relevance to a range of other feminist concerns.

There is, in other words, good reason to think that biological explanations of gender differences will tend to be of little use to feminists even when they are true. Taken together with the insight that the promulgation of such explanations tends to work against feminist

purposes independently of whether they are true or not, these observations provide a powerful rationale for feminists to ignore biological explanations.

6.5 CONCLUSION

In this chapter, I explored potential justifications for a feminist right to ignore biological explanation of gender differences. I argued that Kronfeldner's right to ignore cannot be applied to the feminist case because of its dependence on the assumption of psychic unity. I then considered the frequently invoked "flawed science" and "just ideology" justifications. Although both provide important insights, I argued that they are too limited to establish a general right to ignore. The "flawed science" approach grants the right to ignore biological explanations only on a case-by-case basis. The "just ideology" approach is vulnerable to the objection that biological explanations – even if they are detrimental to feminist aims independently of their truth value – still need to be taken into consideration because they may also harbour important insights for feminist purposes.

Contrary to the discouraging findings so far, the penultimate section suggested that a right to ignore can be defended. The hybrid kind model suggests that social explanations of the status kind *gender* generally have considerable autonomy from biological explanations of the base kind *sex*. This established a limited right to ignore for specific types of feminist explanations that have gender as their explanandum. Taking a different stance on the extant discussion suggested that, in addition to this limited right to ignore, a more far-reaching right to ignore may be within reach. By looking at a range of key feminists debates, and finding that questions of innate cognitive differences are of little or no relevance to any of them, the case for a comprehensive right to ignore could be rehabilitated. Although this discussion did not establish a "come what may" universal right for feminists to ignore biological explanations, it suggests that ignoring such explanations may be a defensible feminist heuristic.

CONCLUSION

My aim in this thesis has been to advance a number of debates on classification and explanation in the social and natural sciences by analysing them from the novel angle of the hybrid kind model. Hybrid kinds, so I argued, are categories that consist of a base kind and an associated status kind. While the base kind is constituted by the properties that are commonly used to identify the category, the status kind is constituted by the social position that individuals acquire *qua* being recognised and treated as members of the specific category.

Although the core idea of this model has been around at least as long as Searle's account of social kinds, its potential as a central tool for understanding the peculiarities of the social world and the social sciences had yet to be unleashed. I argued that the hybrid kind model not only allows us to understand core features of classifications in a social world better than competitors such as Guala's and Epstein's account. Its core virtue consists in making visible a common thread that runs through a number of key debates on the relationship between the natural and the social sciences.

Following this common thread has led me to develop a tentative defence of the view that inquiry into the social world is in important respects different from inquiry into the natural world. Hybrid kinds, which we encounter whenever classifications acquire a meaning ("status") in social contexts, are subject to moral and political considerations in a way non-hybrid kinds are not. The fact that our terminological decisions may have a wide-ranging impact on people's lives suggests that moral and political considerations may legitimately play a direct role in these decisions. In addition, hybrid kinds can pose challenges to scientific inquiry that make them poorly suited to the epistemic role and the scientific approaches associated with natural kinds. These challenges include the tendency towards biased conceptualisation, the diversity and complexity of mechanisms mediating classificatory feedback, and the fact that the social meanings associated with classifications are too diverse and context-specific to make suitable candidates for natural kinds.

Developing these arguments not only demanded establishing a novel understanding of the social world through the lens of the hybrid kind model. It often also required rethinking what we thought we already knew about the natural sciences and natural kinds. Accordingly, I suggested that semantic externalism is untenable not only with regard to the social sciences, but also in its traditional realm of the natural sciences. Furthermore, I argued that natural kinds are not simply vectors for projections and generalisations, but

analytic tools that incorporate assumptions about the causal mechanisms which constitute the kind.

Moving from the topic of classification on to explanation, the case study of evolutionary and social explanations of cognitive gender differences proved an opportunity to argue not only for the difference but also the autonomy of the social sciences from the natural sciences. While the two approaches turned out to be unexpectedly amicable in terms of the pure logical compatibility of their respective causal claims, the picture changed when we considered their methodological and political-pragmatic relevance to each other. On the methodological level, I argued that the empirical evidence used to support social explanations for gender differences tends to undermine certain types of evolutionary (trigger innateness) explanations. This gives rise to an evidential asymmetry between social and biological explanations of gender differences which runs exactly counter to the common pop-science portrayal in which biology calls the shots always. In addition to these methodological considerations, I suggested that there may be pragmatic considerations that make ignoring biological explanations of gender differences a sound heuristic in the context of feminist politics.

BIBLIOGRAPHY

Ahmed, S. (2008). Imaginary prohibitions. Some preliminary remarks on the founding gestures of the 'new materialism'. *European Journal of Women's Studies* 15(1), 23-39.

Anderson, E. (1995). Knowledge, human interests, and objectivity in feminist epistemology. *Philosophical Topics* 23(2), 27-58.

Anderson, E. (2004). Uses of value judgments in science: a general argument, with lessons from a case study of feminist research on divorce. *Hypatia* 19(1), 1-24.

Ariew, A. (1996). Innateness and canalization. *Philosophy of Science* 63(3), S19-S27.

Ariew, A. (1999). Innateness is canalization: in defense of a developmental account of innateness. In V. G. Hardcastle (ed.) *Where biology meets psychology*. Cambridge, MA: MIT Press, 117-138.

Ásta Sveinsdóttir (2008). Essentiality conferred. *Philosophical Studies* 140(1), 135-148.

Ásta Sveinsdóttir (2011). The metaphysics of sex and gender. In C. Witt (ed.) *Feminist metaphysics*, New York: Springer, 47-65.

Ásta Sveinsdóttir (2013). The social construction of human kinds. *Hypatia* 28(4), 716-732.

Ásta Sveinsdóttir (2017). Social kinds. In K. Ludwig & M. Jankovic (eds.) *The Routledge handbook of collective intentionality*. New York: Routledge, 290-299.

Audi, P. (2012a). A clarification and defense of the notion of grounding. In F. Correia & B. Schnieder (eds.) *Metaphysical grounding: understanding the structure of reality*. Cambridge: Cambridge University Press, 101-121.

Audi, P. (2012b). Grounding: toward a theory of the in-virtue-of relation. *Journal of Philosophy* 109, 685-711.

Barkow, J. H. (2006). *Missing the revolution: Darwinism for social scientists*. Oxford: Oxford University Press.

Bashamboo, A. & K. McElreavey (2016). Mechanism of sex determination in humans: insights from disorders of sex development. *Sexual Development* 10, 313-325.

Bateson, P. (1991). Are there principles of behavioural development? In P. Bateson (ed.) *The development and integration of behaviour: essays in honour of Robert Hinde*. Cambridge: Cambridge University Press, 19-39.

Bateson, P. & P. Martin (1999). *Design for a life: how behaviour develops*. London: Jonathan Cape.

Bauman, Z. (1998). *Globalization: The human consequences*. New York: Columbia University Press.

Beal, C. R. (1994). *Boys and girls: the development of gender roles*. New York: McGraw-Hill.

Becker, H. S. (1963). *Outsiders. Studies in the sociology of deviance*. New York: Free Press.

Benedict, R. (1934). *Patterns of culture*. New York: Houghton Mifflin.

Bird, A. (2010). Discovering the essences of natural kinds. In H. Beebee & N. Sabbarton-Leary (eds.) *The semantics and metaphysics of natural kinds*. New York: Routledge, 125-136.

Block, N. J. & G. Dworkin (1976). *The I.Q. controversy: critical readings*. Oxford: Pantheon Books.

Boas, F. (1911). *The mind of primitive man*. New York: Macmillan.

Bogen, J. (1988). Symposium papers, comments and an abstract: comments on 'the sociology of knowledge about child abuse'. *Nous* 22, 65-66.

Bolnick, D. A. (2008). Individual ancestry inference and the reification of race as a biological phenomenon. In B. A. Koenig, S. S. J. Lee & S. S. Richardson (eds.) *Revisiting race in a genomic age*. New Jersey: Rutgers University Press, 70-85.

Boyd, R. (1989). What realism implies and what it does not. *Dialectica* 43, 5-29.

Boyd, R. (1991). Realism, anti-foundationalism, and the enthusiasm for natural Kinds, *Philosophical Studies* 61, 127-148.

Boyd, R. (1999). Kinds, complexity and multiple realization. *Philosophical Studies* 95(1), 67-98.

Boyd, R. & P. Richerson (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Brewer, S. (2001). *A child's world: a unique insight into how children think*. London: Headline.

Brigandt, I. (2009). Natural kinds in evolution and systematics: metaphysical and epistemological considerations. *Acta Biotheoretica* 57, 77-97.

Brigandt, I. (2010). The epistemic goal of a concept: accounting for the rationality of semantic change and variation. *Synthese* 177, 19-40.

Bronstein, C. (2011). *Battling pornography: the American feminist anti-pornography movement 1976–1986*. Cambridge: Cambridge University Press.

Buller, D. J. (2005). *Adapting minds: Evolutionary psychology and the persistent quest for human nature*. Cambridge, MA: MIT Press.

Buss, D. M. (1989a). Sex differences in human mate preferences: Evolutionary hypothesis tested in 37 cultures. *Behavioral and Brain Sciences* 12, 1-49.

Buss, D. M. (1989b). Author's response: Toward and evolutionary psychology of human mating. *Behavioral and Brain Sciences* 12, 39-46.

Buss, D. M. (1990). Evolutionary social psychology: Prospects and pitfalls. *Motivation and Emotion* 14(4), 265-286.

Buss, D. M. (2005). Sex differences in the design features of socially contingent mating adaptations. *Behavioral and Brain Sciences* 28, 278-279.

Buss, D. M. & Barnes, M. (1986). Preferences in human mate selection. *Journal of Personality and Social Psychology* 50(3), 559-570.

Buss, D. M. & Schmitt, D.P. (2011). Evolutionary psychology and feminism. *Sex Roles* 64(9), 768-787.

Butler, J. (1990). *Gender trouble*. New York: Routledge.

Campbell, A. (2002). *A mind of her own: the evolutionary psychology of women*. Oxford: Oxford University Press.

Carr, E. & D. Szymanski (2011). Sexual objectification and substance abuse in young adult women. *The Counseling Psychologist* 39, 39-66.

Cavalli-Sforza, L., Menozzi, P. & A. Piazza (1994). *The history and geography of human genes*. Princeton, NJ: Princeton University

Cassidy, A. (2007). The (sexual) politics of evolution: popular controversy in the late 20th-century United Kingdom. *History of Psychology* 10(2), 199-226.

Chiricos, T., Barrick, K. & W. Bales (2007). The labeling of convicted felons and its consequences for recidivism. *Criminology* 45, 547-581.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Morton.

Chomsky, N. (1966). *Cartesian linguistics: a chapter in the history of rationalist thought*. New York: Harper & Row.

Clarke, C. (2017). Review of Francesco Guala's "Understanding institutions". *The British Journal for the Philosophy of Science, Review of Books*.
<https://bjpsbooks.wordpress.com/2017/05/23/francesco-guala-understanding-institutions/> (last accessed 9 November 2018).

Contratto, S. (2002). A feminist critique of attachment theory and evolutionary psychology. In M. Ballou & L. S. Brown (eds.) *Rethinking mental health and disorder: feminist perspectives*. New York: The Guilford Press, 29-47.

Cooper, R. (2004). Why Hacking is wrong about human kinds. *The British Journal for the Philosophy of Science* 55, 73-85.

Cooper, R. (2005). *Classifying madness: a philosophical examination of the diagnostic and statistical manual of mental disorders*. Dordrecht: Springer.

Cooper, R. (2007). *Psychiatry and philosophy of science*. Stocksfield: Acumen Publishing.

Correia, F. & B. Schnieder (eds.) (2012). *Metaphysical grounding: understanding the structure of reality*. Cambridge: Cambridge University Press.

Craver, C. (2009). Mechanisms and natural kinds. *Philosophical Psychology* 22, 575-94.

Crenshaw, K. W. (1991). Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Review* 43(6), 1241-1299.

Cudd, A. E. (2006). *Analyzing oppression*. New York: Oxford University Press.

Culp, R. E., Cook, A. S. & P. C. Housley (1983). A comparison of observed and reported adult-infant interactions: Effects of perceived sex. *Sex Roles* 9(4), 475-479.

Davis, S. N. (2003). Sex stereotypes in commercials targeted toward children: a content analysis. *Sociological Spectrum* 23(4), 407-424.

Department of Health and Social Care, UK Government (2005). Responding to domestic abuse: a handbook for health professionals.
<https://www.gov.uk/government/publications/domestic-abuse-a-resource-for-health-professionals> (last accessed 9 November 2018).

Devitt, M. & K. Sterelny (1999). *Language and reality: An introduction to philosophy of language*. Cambridge, MA: MIT Press.

Douglas, M. (1986). *How institutions think*. Syracuse, NY: Syracuse University Press.

Drabek, M. (2014). *Classify and label: the unintended marginalization of social groups*, New York: Lexington Books.

Dupre, J. (1981). Natural kinds and biological taxa. *Philosophical Review* 90, 66-91

Dupre, J. (1993). *The disorder of things*. Cambridge. MA: Harvard University Press.

Dupre, J. (2001). *Human nature and the limits of science*. New York: Oxford University Press.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Hillsdale, NJ: Lawrence Erlbaum.

Eccles, J. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly* 11, 135-172.

Epstein, B. (2014). How many kinds of glue hold the social world together? In M. Gallotti & J. Michael (eds.) *Social ontology and social cognition*. Dordrecht: Springer, 41-55.

Epstein B. (2015). *The ant trap: rebuilding the foundations of the social sciences*. New York: Oxford University Press

Epstein, B. (2016). Replies to Guala and Gallotti. *Journal of Social Ontology* 2 (1), 159-172.

Equality and Human Rights Commission (2011). Briefing paper 2: Gender pay gaps. <https://www.equalityhumanrights.com/en/publication-download/briefing-paper-2-gender-pay-gap> (last accessed 9 November 2018).

Ereshefsky, M. & T. Reydon (2015). Scientific kinds. *Philosophical Studies* 172, 969-986.

Fausto-Sterling, A. (1992). *Myths of gender: biological theories about women and men*. New York: Basic Books.

Fausto-Sterling, A. (2000). Beyond difference: feminism and evolutionary psychology. In H. Rose & S. Rose (eds.) *Alas, poor Darwin: arguments against evolutionary psychology*. London: Vintage, 209-227.

Fausto-Sterling, Anne (2000a). *Sexing the body: gender politics and the construction of sexuality*. New York: Basic Books.

Feldman, M. & R. C. Lewontin (2008). Race, ancestry, and medicine. In B. A. Koenig, S. S. J. Lee & S. S. Richardson (eds.) *Revisiting race in a genomic age*. New Jersey: Rutgers University Press, 70-85.

Ferguson, C. & R. Hartley (2009). The pleasure is momentary…the expense damnable? The influence of pornography on rape and sexual assault. *Aggression and Violent Behavior* 14, 323-329.

Fine, C. (2010). *Delusions of gender: the real science behind sex differences*. London: Icon Books.

Fine, C. (2017). *Testosterone Rex: unmaking the myths of our gendered minds*. London: Icon Books.

Fine, K. (2001). The question of realism. *Philosophers' Imprint* 1, 1-30.

Fine, K. (2010). Some puzzles of ground. *Notre Dame Journal of Formal Logic* 51(1), 97-118.

Finlayson, L. (2016). *An introduction to feminism*. Cambridge: Cambridge University Press.

Foster, D. (2016). *Lean out*. London: Repeater Books.

Foucault, M. (1978). *The history of sexuality. Volume 1*. New York: Pantheon Books.

Fox-Keller, E. (2010). *The mirage of a space between nature and nurture*. London and Durham: Duke University Press.

Fracchia, J. & R. C. Lewontin (1999). Does culture evolve? *History and Theory* 38(4), 52-78.

Fredrickson, B. & T. Roberts (1997). Objectification theory: toward understanding women's lived experiences and mental health risks. *Psychology of Women Quarterly* 21, 173-206.

Gangestad, S. W. & J. A. Simpson (2000). The evolution of human mating: trade-offs and strategic pluralism. *Behavioral and Brain Sciences* 23(4), 573-587.

Gannett, L. (2004). The biological reification of race. *The British Journal for the Philosophy of Science* 55(2), 323-345.

Gannett, L. (2005). Group categories in pharmacogenetics research. *Philosophy of Science* 72(5), 1232-1247.

Gannett, L. (2010). Questions asked and unasked: how by worrying less about the 'really real' philosophers of science might better contribute to debates about genetics and race. *Synthese*, 177(3), 363-385.

Germon, J. (2009). *Gender: a genealogy of an idea*. New York: Palgrave Millican.

Geuss, R. (1981). *The Idea of a critical theory*. Cambridge: Cambridge University Press

Goetz, A. T., Shackelford, T. K. & J. A. Camilleri (2008). Proximate and ultimate explanations are required for a comprehensive understanding of partner rape. *Aggression and Violent Behavior* 13(2), 119-123.

Godfrey-Smith, P. (2012) Darwinism and cultural change. *Philosophical Transactions of the Royal Society B* 367, 2160-2170.

Goldfinch, A. (2015). *Rethinking evolutionary psychology*. New York: Palgrave Macmillan.

Gould, S. J. & R. C. Lewontin (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London* 205, 581-598.

Greenwald, A. G. & M. R. Banaji (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review* 102(1), 4-27.

Greenwald, A. G. & L. H. Krieger (2006). Implicit bias: scientific foundations. *California Law Review* 94(4), 945-967.

Griesemer, J. & M. J. Wade (1988). Laboratory models, causal explanation and group selection. *Biology and Philosophy* 3, 67-96.

Griffin, G. (2017). *A dictionary of gender studies*. Oxford: Oxford University Press.

Griffiths, P. E. & E. Machery (2008). Innateness, canalization, and "biologicizing the mind." *Philosophical Psychology* 21(3), 397-414.

Grusec, J. E. & P. D. Hastings (2007). *Handbook of socialization: theory and research*. New York: Guilford Press.

Guala, F. (2010). Infallibilism and human kinds. *Philosophy of the Social Sciences* 40(2), 244-264.

Guala, F. (2014). On the nature of social kinds. In M. Gallotti & J. Michaels (eds.) *Perspectives on social ontology and social cognition*. Dodrecht: Springer, 57-68.

Guala, F. (2016a). Epstein on anchors and grounds. *Journal of Social Ontology* 2(1), 135-147.

Guala, F. (2016b). *Understanding institutions*. Princeton: Princeton University Press.

Hacking, I. (1986). Making up people. In T. C. Heller, M. Sosna, and D. E. Wellbery (eds.) *Reconstructing individualism: autonomy, individuality, and the self in Western thought*. Stanford, CA: Stanford University Press, 222-36.

Hacking, I. (1988). The sociology of knowledge about child abuse. *Nous* 22, 53–63.

Hacking, I. (1991). The making and molding of child abuse. *Critical Inquiry* 17, 253-288.

Hacking, I. (1992). World making by kind making: Child-abuse for example. In M. Douglas & D. Hull (eds.) *How classification works: Nelson Goodman among the social sciences*. Edinburgh: Edinburgh University Press, 180-238.

Hacking, I. (1995a). *Rewriting the soul: multiple personality and the sciences of memory*. Princeton, NJ: Princeton University Press.

Hacking, I. (1995b). The looping effects of human kinds. In D. Sperber, D. Premack, and A. J. Premack (eds.) *Causal cognition: a multidisciplinary debate*. New York: Clarendon Press, 351-394.

Hacking, I. (1997). Taking bad arguments seriously. *London Review of Books* 19, 14-16.

Hacking, I. (1998). *Mad travellers: reflections on the reality of transient mental illnesses*. Charlottesville, VA: University Press of Virginia.

Hacking, I. (1999). *The social construction of what?* Cambridge, MA: Harvard University Press.

Hacking, I. (2006). *Kinds of people: moving targets*. British Academy Lecture, 11 April 2006, <www.britac.ac.uk/pubs/src/_pdf/hacking.pdf> (last accessed 9 November 2018).

Hacking, I. (2007). Natural kinds: rosy dawn, scholastic twilight. *Royal Institute of Philosophy Supplements* 61, 203-239.

Häggqvist, S. & A. Wikforss (2017). Natural kinds and natural kind terms: myth and reality. *The British Journal for the Philosophy of Science*, online first. <https://doi.org/10.1093/bjps/axw041> (last accessed 9 November 2018).

Hargreaves Heap, S. & Y. Varoufakis (1995). *Game theory: a critical introduction*. London: Routledge.

Haslanger, S. (2012): *Resisting reality: social construction and social critique*. New York: Oxford University Press.

Haslanger, S. (2014a). Theorizing with a purpose: the many kinds of sex. In M. Gallotti & J. Michael (eds.) *Social ontology and social cognition*. Dordrecht: Springer.

Haslanger, S. (2014b). Race, intersectionality, and method: a reply to critics. *Philosophical Studies* 171, 109-119.

Haslanger, S. (2015). Distinguished lecture: social structure, narrative and explanation. *Canadian Journal of Philosophy* 45(1), 1-15.

Haslanger, S. (2016). What is a (social) structural explanation? *Philosophical Studies* 173(1), 113-130.

Hawley, K. (2017). Comments on Brian Epstein's The Ant Trap. *Inquiry*, online first.

Hendry, R. F. (2006). Elements, compounds, and other chemical kinds. *Philosophy of Science* 73, 864-875.

Heyes, C. (2011). What's social about social learning? *Journal of Comparative Psychology* 126(2), 193-202.

Heyes, C. (2012). Grist and mills: on the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society B* 367, 2181–2191.

Heyes, C. (2018). *Cognitive gadgets: the cultural evolution of thinking*. Cambridge, MA: Harvard University Press.

Jablonka, E. & M. J. Lamb (2005). *Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life*. Cambridge, MA: MIT Press.

Jackson, F. & P. Pettit (1992). Structural explanation in social theory. In K. Lennon & D. Charles (eds.) *Reduction, explanation, and realism*. Oxford: Oxford University Press, 97-131.

Jackson, J. P. (2010). Definitional argument in evolutionary psychology and cultural anthropology. *Science in Context* 23, 121-150.

Jenkins, K. (2016). Amelioration and inclusion: gender identity and the concept of woman. *Ethics* 126 (2), 394-421.

Kay, A. & A. Furnham (2013). Age and sex stereotypes in British television advertisements. *Psychology of Popular Media Culture* 2(3), 171-186.

Kaplan, J. M. (2010). When socially determined categories make biological realities: understanding Black/White health disparities in the U.S. *The Monist* 93, 281-297.

Kaplan, J. & R. Winther (2013). Prisoners of abstraction? The theory and measure of genetic variation, and the very concept of "race". *Biological Theory* 7, 401-412.

Khalidi, M. A. (2001). Innateness and domain specificity. *Philosophical Studies* 105(2), 191-210.

Khalidi, M. A. (2010a). Interactive kinds. *The British Journal for the Philosophy of Science* 61, 335-60.

Khalidi, M. A. (2010b). What is domain specificity (and why does it matter)? In S. Ohlsson & R. Catrambone (eds.) *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin TX: Cognitive Science Society, 194-199.

Khalidi, M. A. (2013). *Natural categories and human kinds: classification in the natural and social sciences*. New York: Cambridge University Press.

Khalidi, M. A. (2015). Three kinds of social kinds. *Philosophy and Phenomenological Research* 90(1), 96-112.

Kitcher, P. (1984). Species. *Philosophy of Science* 51, 308-333.

Kitcher, P. (2011). *Science in a democratic society*. New York: Prometheus.

Kripke, S. (1980). *Naming and necessity*. Harvard: Harvard University Press.

Kroeber, A. (1917). The superorganic. *American Anthropologist* 19, 163-213.

Kronfeldner, M. (2009). If there is nothing beyond the organic... *NTM - Journal of the History of Science, Technology and Medicine* 17, 107-133.

Kronfeldner, M. (2010). Darwinian 'blind' hypothesis formation revisited. *Synthese* 175(2), 193-218.

Kronfeldner, M. (2011). *Darwinian creativity and memetics*. London: Routledge.

Kronfeldner, M. (2017). The right to ignore: an epistemic defence of the nature/culture divide. In R. Joyce (ed.) *The Routledge handbook of evolution and philosophy*. New York: Routledge, 210-224.

Kronfeldner M. (2018). *What's left of human nature? A post-essentialist, pluralist and interactive account of a contested concept*. Cambridge, MA: MIT Press.

Kuorikoski, J. & S. Pöyhönen (2012). Looping kinds and social mechanisms. *Sociological Theory* 30, 187-205.

Laimann, J. (forthcoming). Capricious Kinds. *The British Journal for the Philosophy of* Science, first online 14 March 2018. doi:10.1093/bjps/axy024.

Laland, K. (2004). Social learning strategies. *Animal Learning & Behavior* 32(1), 4-14.

Laland, K., Uller, T., Feldman, M., Sterelny, K., Müller, G., Moczek, A., Jablonka, E. & J. Odling-Smee (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings of the Royal Society B: Biological Sciences* 282(1813), 20151019.

LaPorte, J. (2010). Theoretical identity statements, their truth, and their discovery. In H. Beebee & N. Sabbarton-Leary (eds.*) The semantics and metaphysics of natural kinds*. New York: Routledge, 115-124.

Lauzen, M. M., Dozier, D. M. & N. Horan (2008). Constructing gender stereotypes through social roles in prime-time television. *Journal of Broadcasting & Electronic Media* 52(2), 200-214.

Lemert, E. (1951). *Social pathology*. New York: McGraw-Hill.

Lewens, T. (2003). Prospects for evolutionary policy. *Philosophy* 78, 483-502.

Lewens, T. (2015). *Cultural evolution: conceptual challenges*. Oxford: Oxford University Press.

Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

Lewontin, R. (1972). The apportionment of human diversity. *Evolutionary Biology* 6 (1972): 381-398.

Lewontin, R. C., Rose, S. P. R. & L. J. Kamin (1984). *Not in our genes: biology, ideology, and human nature*. New York: Pantheon Books.

Liesen, L. T. (2007). Women, behavior, and evolution. *Politics and the Life Sciences* 26(1), 51-70.

Lloyd, E. A. (1988). *The structure and confirmation of evolutionary theory.* New York: Greenwood Press.

Lloyd, E. A. (2001). Units and levels of selection: an anatomy of the units of selection debates. In R. S. Singh et al. (eds.) *Thinking about evolution*. Cambridge: Cambridge University Press, 267-291.

Longino, H. (1990). *Science as social knowledge: values and objectivity in scientific inquiry.* Princeton: Princeton University Press.

Longino, H. (2013). *Studying human behavior: how scientists investigate aggression and sexuality.* Chicago: University of Chicago Press.

Lucal, B. (2010). Better informed, still skeptical: response to Machalek and Martin. *Teaching Sociology* 38(1), 46-49.

Ludwig, D. (2017). Ontological choices and the value-free ideal. *Erkenntnis* 81, 1253-1272.

Malamuth, N., T. Addison & M. Koss (2000). Pornography and sexual aggression: are there reliable effects and can we understand them? *Annual Review of Sex Research* 11. 26-91.

MacKinnon, C. (1987). *Feminism unmodified: discourses on life and law*. Cambridge, MA: Harvard University Press.

MacKinnon, C. (1989). *Toward a feminist theory of the state*. Cambridge, MA: Harvard University Press.

Magnus, P. D. (2014). NK ≠ HPC. *Philosophical Quarterly* 64 (256), 471-77.

Mallon, R. (2016). *The construction of human kinds*, Oxford: Oxford University Press.

Martin, B. (2008). Success and its failures. In J. W. Scott (ed.) *Women's studies on the edge.* London and Durham: Duke University Press, 169-197.

Mason, R. (2016). The metaphysics of social kinds. *Philosophy Compass* 11, 841-850.

Mayr, E. (1963). *Animal Species and Evolution.* Cambridge, MA: Harvard University Press.

McKibbin, W. F., Shackelford, T. K. & A. T. Goetz (2008). Why do men rape? An evolutionary psychological perspective. *Review of General Psychology* 12(1), 86-97.

McMullin, E., (1982). Values in Science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 3-28.

McNeil, M. C., Polloway, E. A. & J. D. Smith (1984). Feral and isolated children: historical review and analysis. *Education and training of the mentally retarded* 19(1), 70-79.

Mead, M. (1928). *Coming of age in Samoa*. New York: William Morrow and Co.

Mead, M. (1932*). Sex and temperament in three primitive societies*. New York: William Morrow and Co.

Mead, M. (1949). *Male and female: the classic study of the sexes*. New York: William Morrow and Co.

Meloni, M. (2016a). *Political biology. Science and social values from eugenics to epigenetics*. London: Palgrave.

Meloni, M. (2016b). From boundary-work to boundary object: how biology left and re-entered the social sciences. In M. Meloni, S. Williams & P. Martin (eds.) *Biosocial matters: rethinking the sociology-biology relations in the twenty-first century*. New York: Wiley-Blackwell.

Mesoudi A., Whiten A., K. N. Laland (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences* 29, 329-383.

Meyer, J. W. & B. I. Sobieszek (1972). Effect of a child's sex on adult interpretations of its behavior. *Developmental Psychology* 6(1), 42-48.

Mikkola, M. (2017). Grounding and anchoring: on the structure of Epstein's social ontology. *Inquiry*, online first.

Mikkola, M. (2006). Elizabeth Spelman, gender realism, and women. *Hypatia* 21, 77-96.

Mikkola, M. (2017). Feminist perspectives on sex and gender. *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), E. N. Zalta (ed.).

<https://plato.stanford.edu/archives/fall2017/entries/feminism-gender/> (last accessed 9 November 2018).

Mill, J. S. (1843). *A system of logic*. London: J. W. Parker.

Mill, J. S. (1984). The subjection of women, Essays on equality, law, and education. In J. Robson (ed.) *Collected works of John Stuart Mill* vol. XXI. Toronto: Toronto University Press, 259–348. (Original printed in 1869).

Millikan, R. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, MA: MIT Press.

Millikan, R. (1999). Historical kinds and the 'special sciences'. *Philosophical Studies* 95, 45-65.

Mills, C. (2005). "Ideal theory" as ideology. *Hypatia* 20(3), 165-184.

Money, J. (1955). Hermaphroditism, gender and precocity in hyper-adrenocorticism: psychological findings. *Bulletin of Johns Hopkins Hospital* 96(3), 253-263.

Money, J. (1957). *The psychological study of man*. Springfield, Ill.: Charles C. Thomas.

Moradi, B. & Y. Huang (2008). Objectification theory and psychology of women: a decade of advances and future directions. *Psychology of Women Quarterly* 32, 277-398.

Murphy, D. (2006). *Psychiatry in the scientific image*. Cambridge, MA: MIT Press.

Newcombe, N. S. (2010). On tending to our scientific knitting: thinking about gender in the context of evolution. In J. C. Chrisler & D. R. McCreary (eds.) *Handbook of gender research in psychology,* Volume 1. New York: Springer, 259-274.

O'Donnell, A.T., Corrigan, F., & S. Gallagher (2015). The impact of anticipated stigma on psychological and physical health problems in the unemployment group. *Frontiers in Psychology* 6, 1263.

O'Malley, M. A. (2014). *Philosophy of microbiology*. Cambridge: Cambridge University Press.

Oakley, A. (1972). *Sex, gender and society*. London: Temple Smith.

Office for National Statistics (2015). *Statistical bulletin: annual Survey of hours and earnings: 2015 Provisional Results*.

<https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2015provisionalresults> (last accessed 9 November 2018).

Okasha, S. (2006). *Evolution and levels of selection*. Oxford: Oxford University Press.

Okin, S. M. (1989). *Justice, gender, and the family*. New York: Basic Books.

Oyama, S. (2000). *Evolution's eye: a systems view of the biology-culture divide*. Durham and London: Duke University Press.

Pinker, S. (2002*). The blank slate: the modern denial of human nature*. London: Penguin.

Pray, L. (2008). Antibiotic resistance, mutation rates and MRSA'. *Nature Education* 1, 30.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7, 215-271.

Radcliffe-Richards, J. (1998). Feminism and equality. *Journal of Contemporary Legal Studies* 9, 225-247.

Radcliffe-Richards, J. (2014). Only X%: the problem of sex equality. *Journal of Practical Ethics* 2(1), 44-67.

Ray, E. & C. Heyes (2011). Imitation in infancy: the wealth of the stimulus. *Developmental Science* 14(1), 92-105.

Reydon, T. (2009). How to fix kind membership: a problem for HPC theory and a solution, *Philosophy of Science* 76, 724-736.

Richerson, P. & R. Boyd (2005). *Not by genes alone: how culture transformed human evolution.* Chicago: University of Chicago Press.

Ridgeway, C. (2011). *Framed by gender: how gender inequality persists in the modern world*. Oxford: Oxford University Press.

Rooney, P. (1992). On values in science: is the epistemic/non-epistemic distinction useful? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 13-22.

Rosen, G. (2010). Metaphysical dependence: grounding and reduction. In R. Hale & A. Hoffman (eds.) *Modality: metaphysics, logic, and epistemology*. Oxford: Oxford University Press, 109-136.

Rosen, G. (2015). Real definition. *Analytic Philosophy* 56(3), 189-209.

Rudman, L. A. & P. Glick (2010). *The social psychology of gender: how power and intimacy shape gender relations*. New York: Guilford Press.

Saul, J. (2006). Gender and race. Philosophical analysis and social kinds. *Proceedings of the Aristotelian Society* 80, 119-143.

Sax, L. (2002). How common is intersex? A response to Anne Fausto-Sterling. *The Journal of Sex Research* 39(3), 174-178.

Schaffer, J. (forthcoming). Anchoring as grounding: On Epstein's The Ant Trap. *Philosophy and Phenomenological Research*.

Schaffer, J. (2009). On what grounds what. In D. Chalmers et al. (eds.) *Metametaphysics*. Oxford: Oxford University Press, 347-383.

Scharpf, C. (2000). Politics, science, and the fate of the Alabama sturgeon. *American Current* 26(3), 6-14.

Schmitt, D. P. (2005). Sociosexuality from Argentina to Zimbabwe: a 48-nation study of sex, culture, and strategies of human mating. *Behavioral and Brain Sciences* 28(2), 247-275.

Schur, E. M. (1971). *Labeling deviant behavior*. New York: Harper Row.

Scully, D. (1990). *Understanding sexual violence: a study of convicted rapists*. London: HarperCollins Academic.

Searle J. (1995). *The construction of social reality*. New York: The Free Press.

Searle, J. (2010). *Making the social world*. Oxford: Oxford University Press.

Segal, L. (2000). Gender, genes and genetics: from Darwin to the human genome. In C. Squire (ed.) *Culture and psychology*. London: Routledge, 31-43.

Segerstråle, U. C. O. (2000). *Defenders of the truth: the battle for science in the sociobiology debate and beyond*. Oxford: Oxford University Press.

Shea, N. (2012). New thinking, innateness and inherited representation. *Philosophical Transactions of the Royal Society B* 367(1599), 2234-2244.

Slater, M. (2015). Natural kindness. *The British Journal for the Philosophy of Science* 66, 375-411.

Smith, L. J. (1994). A content analysis of gender differences in children's advertising. *Journal of Broadcasting & Electronic Media* 38(3), 323-337.

Sober, E. & D. S. Wilson (1998). *Unto others*. Cambridge, MA: Harvard University Press.

Sobieszek, B. I. (1978). Adult interpretations of child behavior. *Sex Roles* 4(4), 579-588.

Spelman, E. (1988). *Inessential Woman*. Boston: Beacon Press.

Spencer, S., Steele, C. & D. Quinn (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology* 35, 4-28.

Stanford, K. & P. Kitcher (2000). Refining the causal theory of reference for natural kind terms. *Philosophical Studies* 97, 99-129.

Sterelny, K. (2005). Made by each other: organisms and their environment. *Biology & Philosophy* 20(1), 21-36.

Stich, S. P. (1975). Introduction: the idea of innateness. In S. P. Stich (ed.) *Innate Ideas*. Berkeley, CA: University of California Press.

Stoljar, N. (2011). Different women. Gender and the realism-nominalism debate. In C. Witt (ed.) *Feminist Metaphysics*. Dordrecht: Springer.

Symons, D. (1989). The psychology of human mate preferences. *Behavioral and Brain Sciences* 12, 34-35.

Tabery, J. (2014). *Beyond versus: the struggle to understand the interaction of nature and nurture*. Cambridge, MA: MIT Press.

Taylor, P. J. (2014). *Nature-nurture? No: moving the sciences of variation and heredity beyond the gaps*. Cambridge, MA: The Pumping Station.

Thomasson, A. (2003). Realism and human kinds. *Philosophy and Phenomenological Research* 67, 580-609.

Thornhill, R. & C. T. Palmer (2000*). A natural history of rape: biological bases of sexual coercion*. Cambridge, MA: MIT Press.

Tierney, J. (2016). The real war on science. *The City Journal* Autumn 2016. New York: New York Manhattan Institute. <https://www.city-journal.org/html/real-war-science-14782.html> (last accessed 9 November 2018).

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20, 410-433.

Tooby, J. & L. Cosmides (1989a). Adaption versus phylogeny: the role of animal psychology in the study of human behavior. *International Journal of Comparative Psychology* 2(3), 175-188.

Tooby, J. & L. Cosmides (1989b). The innate versus the manifest: How universal does universal have to be? *Behavioral and Brain Sciences* 12(1), 36-37.

Tooby, J. & L. Cosmides (1990). On the universality of human nature and the uniqueness of the individual: the role of genetics and adaptation. *Journal of Personality* 58(1), 17-67.

Tooby, J. & L. Cosmides (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (eds.) *The adapted mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press, 19-136.

Tooby, J. & L. Cosmides (2005). Conceptual foundations of evolutionary psychology. In D. Buss (ed.) *The Handbook of Evolutionary Psychology*. New Jersey: Wiley, 5-67.

Troisi, A. (2001). Gender differences in vulnerability to social stress. *Physiology & Behavior*, 73(3), 443-449.

Tylor, E. B. (1871). *Primitive culture. Researches into the development of mythology, philosophy, religion, language, art and custom*. London: Murray.

Waynforth, D. & R. Dunbar (1995). Conditional mate choice strategies in humans: evidence from "Lonely Hearts" advertisements. *Behaviour* 132(9), 755-779.

Weisgram, E. & L. Dinella (eds.) (2018). *Gender typing of children's toys: how early play experiences impact development*. Washington, DC: APA Books.

West, C. & D. Zimmerman (1987). Doing gender. *Gender and Society* 1, 125-51.

Worrall, J. L. & R. Morris (2011). Inmate custody levels and prison rule violations. *The Prison Journal* 91, 131-57.